

München. 1960

Komponentenanalyse

Theoretische Grundlagen, Rechenverfahren
und einige Anwendungen
in den Wirtschafts- und Sozialwissenschaften

von

AXEL-RAINER WULSTEN



Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41555220_0002

K O M P O N E N T E N A N A L Y S E

Theoretische Grundlagen, Rechenverfahren
und einige Anwendungen in den
Wirtschafts- und Sozialwissenschaften

Inaugural - Dissertation
zur Erlangung der Doktorwürde
einer Hohen Staatswirtschaftlichen Fakultät
der Ludwig-Maximilians-Universität
zu München

vorgelegt von

Axel-Rainer Wulsten

1 9 6 0



Referent: Professor Dr. Hans Kellerer
Korreferent: Professor Dr. Heinrich Strecker
Tag der mündlichen Prüfung: 13. Juli 1960

VORWORT

Die vorliegende Dissertation ist während meiner mehr-jährigen Tätigkeit am Institut für Statistik und ihre Anwendungen in den Wirtschafts- und Sozialwissenschaften der Universität München entstanden. Für stets wohlwollende Unterstützung sowie viele wertvolle Hinweise und Ratschläge möchte ich daher ganz besonders dem Direktor des Institutes, meinem Referenten und akademischen Lehrer, Herrn Professor Dr. Hans K e l l e r e r , danken. Herrn Professor Dr. Heinrich S t r e c k e r , Tübingen, bin ich für sein ständiges Interesse am Fortgang der Arbeit und zahlreiche Anregungen, insbesondere zur Gestaltung des theoretischen Teiles, zu großem Dank verpflichtet. Ursprüngliche Anregung sowie erste Schritte auf dem speziellen Fachgebiet der Arbeit verdanke ich Herrn Professor Dr. Wolfgang W e t z e l , Kiel.

München, im Mai 1960

A.-R. W.

INHALTSVERZEICHNIS

	<u>Seite</u>
0 <u>EINLEITUNG</u>	VII
1 <u>THEORIE DER KOMPONENTENANALYSE</u>	
10 HISTORISCHE VORBEMERKUNG	1
11 PROBLEMSTELLUNG UND ELEMENTARE VORAUSSETZUNGEN	
110 Ausgangsgesamtheit; Aufgabe der Komponentenanalyse	2
111 Normierung der Variablen	3
112 Die Komponenten und ihre Beziehungen zu den Ausgangsvariablen	5
12 KOMPONENTENANALYSE UND LINEARE ALGEBRA: VEKTOREN UND VEKTORRÄUME	
120 Die Variablen als Vektoren	6
121 Vektorraum und euklidischer Raum	9
122 Die Komponentenanalyse als Problem der linearen Algebra	12
13 KOMPONENTENANALYSE UND LINEARE ALGEBRA: MATRIZEN	
130 Definition der vorkommenden Matrizen	15
131 Dimension und Rang des Komponentenproblems	19
14 DETERMINATION DES PROBLEMS DURCH HAUPTKOMPONENTEN	
140 Die Einführung von Hauptkomponenten	20
141 Hauptkomponenten und Varianzen der Ausgangsvariablen	23
142 Vorbereitung der Extremwertaufgabe: Nebenbedingungen	25
143 Lösung der Extremwertaufgabe	26
144 Ergebnis der Extremwertaufgabe	29

15	DIE HAUPTKOMPONENTENLÖSUNG ALS EIGENWERT- PROBLEM DER KORRELATIONSMATRIX	
150	Überführung in das Eigenwertproblem (Erste Hauptkomponente)	31
151	Die Eigenwerte der Korrelationsmatrix . .	33
152	Eigenvektoren, mehrfache Eigenwerte und System der Eigenachsen	37
16	DURCHFÜHRUNG DER HAUPTKOMPONENTENLÖSUNG	
160	Fortführung des Eigenwertproblems (Zweite Hauptkomponente)	42
161	Identifizierung von Eigenwerten und Eigenvektoren nachgeordneter Residual- matrizen; Geometrie der Hauptkomponenten- lösung	46
162	Allgemeine Lösung (k-te Hauptkomponente).	48
163	Gesamtvariation und Eigenwerte der Korrelationsmatrix	51
17	DIE ERGEBNISSE DER HAUPTKOMPONENTENLÖSUNG UND IHRE BEZIEHUNG ZUR HAUPTACHSENTTRANSFORMATION	
170	Gewinnung der Hauptkomponenten	52
171	Nachweis der Eigenschaften der Hauptkomponenten	54
172	Komponentenanalyse und Hauptachsen- transformation	55
173	Spezielle Deutung der Hauptachsen- transformation	58

2 NUMERISCHE KOMPONENTENANALYSE

20	DIE RECHENTECHNIK DER KOMPONENTENANALYSE	
200	Allgemeines zur Rechentechnik	61
201	Abschätzung der Eigenwerte; direkte Verfahren zur Eigenwertbestimmung	63
202	Indirekte Eigenwertbestimmung: Iterationsverfahren	67
203	Ergänzungen zur Rechentechnik: Iterationsverfahren nach HOTELLING, Methode von TUCKER	69

21 EIN ZAHLENBEISPIEL

210	Ausgangsgesamtheit und Korrelationsmatrix	73
211	Die Berechnung der Eigenwerte	75
212	Eigenvektoren und Iterationsprozeß	78
213	Die Hauptkomponenten; Interpretation der Ergebnisse	80

3 ENTWICKLUNG DER KOMPONENTENANALYSE UND
ABGRENZUNG GEGENÜBER DER FAKTORANALYSE

30 EINIGE BEITRÄGE ZUR ENTWICKLUNG DER
KOMPONENTENANALYSE

300	Vorbemerkung	84
301	Die "Geraden bester Anpassung" von KARL PEARSON und ihre Deutung durch KENDALL	85
302	Die Komponentenanalyse nach HOTELLING, GIRSHICK und VAN DEN BOGAARD-KAPTEIN	91
303	Komponentenanalyse und mehrdimensionale Analyse: WILKS, T.W.ANDERSON und TINTNER	97

31 KOMPONENTENANALYSE UND FAKTORANALYSE

310	Kurzer Abriß der multiplen Faktoranalyse	101
311	Über die Unterscheidung zwischen Komponentenanalyse und Faktoranalyse	115

4 ANWENDUNGEN DER KOMPONENTENANALYSE IN DEN
WIRTSCHAFTS- UND SOZIALWISSENSCHAFTEN

40 ALLGEMEINE ANWENDUNGSPROBLEME; METHODISCHE
EINGLIEDERUNG DER KOMPONENTENANALYSE

400	Überblick der Anwendungsgebiete	122
401	Die Stellung der Komponentenanalyse im Rahmen anderer statistischer Verfahren	125

41	ÖKONOMETRIE UND INDEXZAHLEN	
410	Die Hauptkomponenten volkswirtschaftlicher Bewegungsvorgänge: Eine Untersuchung von STONE	129
411	Identifizierung der Hauptkomponenten mit ökonomischen Variablen	137
412	Einige Anwendungen auf Indexzahlen .	140
413	Die Komponentenanalyse als Hilfsmittel zur Indexkonstruktion	148
414	Kombinationen von Indexzahlen	156
415	Anwendungen auf Wirtschaftsrechnungen: Untersuchung der Verbrauchsstruktur .	160

42	SPEZIELLE ANWENDUNGSGEBIETE	
420	Die Verwendung von Hauptkomponenten in der Straßenverkehrsunfallstatistik	164
421	Komponentenanalyse und statistische Regionalforschung	170
422	Anwendungen in der technischen Statistik; Ausblick auf verwandte Probleme	178

VERZEICHNIS DER ABKÜRZUNGEN FÜR ZEITSCHRIFTEN	183
---	-----

LITERATURVERZEICHNIS	185
----------------------	-----

0 EINLEITUNG

Bei der Anfertigung dieser Arbeit hat sich der Verfasser ein doppeltes Ziel gesetzt. Einerseits besteht die Absicht, eine zur Analyse mehrdimensionaler Gesamtheiten von quantitativen Daten entwickelte Methode der mathematischen Statistik ausführlich darzustellen; andererseits soll versucht werden, die Anwendungsmöglichkeiten dieser Methode in den Wirtschafts- und Sozialwissenschaften in einer möglichst umfassenden Weise zu behandeln. Daß somit zwei Zwecke hervorgehoben werden, obwohl der Gesamtaufbau der Arbeit keinen anderen als den gewöhnlichen Weg von der theoretischen Grundlegung eines statistischen Verfahrens zu seiner praktischen Anwendung erkennen läßt, hat folgende Ursachen.

Die im theoretischen Teil herzuleitende Methode findet im Rahmen der üblichen Begriffsbestimmung ihren Standort innerhalb der Faktoranalyse oder doch in ihrer Nähe, daher sah der Verfasser seine Aufgabe ursprünglich darin, die Faktoranalyse als Kern seiner Ausführungen aufzufassen und die wirtschaftswissenschaftlichen Anwendungen in entsprechender Orientierung anzuschließen. Aus der Beschäftigung mit der Faktoranalyse erwuchs jedoch die Überzeugung, daß unter deren verschiedenen Formen für den vorgesehenen Zweck aus mehreren Gründen - die im Zuge der Darstellung näher zu erläutern sind - eine methodische Variante den Vorzug verdient; es schien daher geboten, die theoretischen Erörterungen auf diese zu beschränken und erst später eine Verbindung zur Faktoranalyse herzustellen. Neuere Bestrebungen zur Verselbständigung dieses speziellen Verfahrens unter dem Namen "Komponentenanalyse" (component analysis) ermöglichten auch die äußerliche Kennzeichnung dieses Vorgehens. Sind nun deutschsprachige Abhandlungen über die Faktoranalyse nicht häufig und überdies ganz überwiegend auf ihr eigentliches Anwendungsgebiet, die Psychologie, ausgerichtet, so gilt für die Komponentenanalyse im besonderen, daß einführende Texte in deutscher Sprache im Rahmen der in

dieser Arbeit vorgestellten Weise bisher vermutlich fehlen. Der Verfasser hat aus dieser freilich nur auf seinen eigenen Überblick gegründeten Vermutung die Berechtigung abgeleitet, ein zunächst auf keinerlei Anwendung bezogenes System der Komponentenanalyse vollständig zu entwickeln und ausführlich darzustellen. Es bedarf keiner Erwähnung, daß die entsprechenden Arbeiten meist amerikanischer oder englischer Statistiker und Psychologen das Vorbild für eine im übrigen durchaus selbständige Herleitung bildeten, ebenso wie Theoreme der linearen Algebra benutzt wurden, die zu den elementaren Ergebnissen der höheren Mathematik gehören. Gerade in letztgenannter Hinsicht ist besonderer Wert auf eine verhältnismäßig breite Form der Darstellung gelegt worden, so daß nur einige Grundsätze des Rechnens mit Vektoren und Matrizen als Voraussetzung gelten, speziell aber das Eigenwertproblem eingehende Beachtung findet.

Der praktische Teil der Arbeit findet seine wesentliche Begründung darin, daß die Komponentenanalyse sich bisher mit wenigen Ausnahmen dem Verwendungsbereich ihrer psychologisch-statistischen Herkunft nicht entzogen hat. Sie ist aber unter den Verfahren der Faktoranalyse dasjenige, dessen methodische Einzelheiten am wenigsten implizierte Bestandteile psychologischer Anwendung aufweisen und das sich deshalb ganz besonders für andere Zwecke eignet. Daraus folgt für diese Arbeit die Aufgabe, die im Rahmen der Wirtschafts- und Sozialwissenschaften bereits vorgenommenen Anwendungen der Komponentenanalyse - soweit methodisch übertragbar auch der Faktoranalyse - zu sammeln und im Hinblick auf die Ausführungen des theoretischen Teiles zu würdigen, ferner den Nutzen von zusätzlichen Anwendungsmöglichkeiten zu untersuchen.

In der Gliederung der Arbeit ist auf eine strenge Zweiteilung verzichtet worden, weil einige Abschnitte ihrem Inhalt nach als zu beiden Bereichen gehörig betrachtet werden können. Es bestehen somit vier größere Abschnitte, von denen

der erste und vierte den Kern der theoretischen beziehungsweise praktischen Ausführungen enthalten, während der zweite und dritte den zwischen diesen beiden Hauptgesichtspunkten erforderlichen Übergang bilden. Im Abschnitt 1 wird in der bereits begründeten Ausführlichkeit eine in sich geschlossene Theorie der Komponentenanalyse dargeboten. Der Abschnitt 2 beschäftigt sich in theoretischer und praktischer Form mit Problemen, die die numerische Handhabung der Komponentenanalyse betreffen. Es werden mehrere Rechenverfahren behandelt und mit Hilfe eines einfachen Zahlenbeispiels konkret erläutert. Für den Abschnitt 3 ist neben einem Überblick der verschiedenen Ausprägungen, die die Komponentenanalyse im Laufe ihrer Entwicklung erfahren hat, eine kurze Zusammenfassung der Faktoranalyse vorgesehen, an die sich die für den gesamten Aufbau der Arbeit wichtige Gegenüberstellung von Komponentenanalyse und Faktoranalyse anschließt. Am Ende steht schließlich der Abschnitt 4 mit einer eingehenden Erörterung der Anwendungen der Komponentenanalyse in den Wirtschafts- und Sozialwissenschaften. Nicht in jedem Falle ist es möglich gewesen, die in diesem Abschnitt enthaltenen Anregungen für andere als die schon einmal durchgeführten Verwendungen in ihrer ganzen Tragweite zu überprüfen, auch ist sicher, daß der Bereich der diesbezüglichen Möglichkeiten keineswegs vollständig abgegrenzt wurde. Der Verfasser möchte seine Ziele als erreicht ansehen, wenn diese Arbeit als ein Beitrag zu weiteren Forschungen in jener Richtung gewertet werden kann.

Entwicklung und wesentliche Züge der heutigen Gestalt der Komponentenanalyse sind auf eine grundlegende Arbeit von HOTELLING ¹⁾ zurückzuführen, die 1933 als ein Beitrag zu der in dieser Zeit lebhaften Diskussion über bestimmte Möglichkeiten quantitativer Forschung in der Psychologie - nämlich über die verschiedenen Modellvorstellungen und Methoden der Faktoranalyse - erschien. In ihren mathematischen Grundzügen war die Komponentenanalyse bereits um die Jahrhundertwende durch K. PEARSON ²⁾ formuliert worden, jedoch beschrieb HOTELLING ein Verfahren, das er unabhängig von PEARSON entwickelt hatte und das innerhalb der Faktoranalyse einen neuartigen Weg wissenschaftlichen Vorgehens eröffnete. Die Komponentenanalyse wurde von HOTELLING selbst sowie von GIRSHICK weiter ausgebildet und in neuerer Zeit ohne direkte Bezugnahme auf spezielle Anwendungen von WILKS, KENDALL, T.W. ANDERSON, VAN DEN BOGAARD und KAPTEIN als Zweig der mehrdimensionalen statistischen Analyse (multivariate analysis) behandelt ³⁾. Von Psychologen wird sie als Methode der Faktoranalyse in den Standardwerken dieses Gebietes beschrieben, eine Reihe namhafter Forscher stellt

-
- 1) HOTELLING, H. (1933) - "Analysis of a complex of statistical variables into principal components", J. Educ. Psychol. 24, Seiten 417-441 und 498-520.
 - 2) PEARSON, K. (1901) - "On lines and planes of closest fit to systems of points in space", Phil. Mag. 6th Ser. 2, Seite 559-572.
 - 3) Es darf bezweifelt werden, ob der deutsche Ausdruck "mehrdimensionale Analyse" den Inhalt des Fachworts "multivariate analysis" völlig zutreffend wiedergibt. Wir folgen jedoch grundsätzlich den Übersetzungsvorschriften des deutsch-englischen Teils des offiziellen Wörterbuchs des Internationalen Statistischen Institutes, vgl. KENDALL, M.G.; BUCKLAND, W.R. (1957) - A Dictionary of Statistical Terms, London (Edinburgh): Oliver & Boyd, Seite 384.

ihre Anwendung in anderen Disziplinen dar, so innerhalb der Wirtschaftswissenschaften vor allem TINTNER und STONE ⁴⁾.

Im folgenden Abschnitt soll die Komponentenanalyse unter Berücksichtigung der Ergebnisse der genannten Verfasser entwickelt werden, wobei den mathematischen Grundlagen stärkere Beachtung zuteil wird. Wir gehen nun zur Ableitung der Methode über.

11 PROBLEMSTELLUNG UND ELEMENTARE VORAUSSETZUNGEN

110 Ausgangsgesamtheit: Aufgabe der Komponentenanalyse

Gegeben sei eine Reihe von n Variablen und eine Anzahl N von Elementen einer statistischen Masse ($n \approx N$). Die Variablen sollen als numerische Ergebnisse (Merkmalsausprägungen) von Beobachtungen aufgefaßt werden, die an den Elementen vorgenommen wurden. Wir wählen das Symbol X für die einzelnen Beobachtungswerte und wollen unter X_{ij} ($i=1,2,\dots,n$; $j=1,2,\dots,N$) den für das Element j beobachteten Wert der Variablen i verstehen.

Wir verfügen über insgesamt (nN) solcher Beobachtungswerte. Ihre sachliche Bedeutung ist für die Ausführungen dieses Abschnitts vorerst von geringerem Interesse. Man kann zum Beispiel annehmen, daß es sich um N Verwaltungsbezirke - etwa Landkreise oder vergleichbare Gemeinden - mit je n Merkmalen handelt. Wir können uns auch n psychologische Tests vorstellen, denen jede einzelne von insgesamt N Personen unterworfen wurde, so daß die Ausgangsgesamtheit aus (nN) Testergebnissen - in beliebiger Weise als quantifiziert unterstellt - besteht.

4) Mit den bisher aufgeführten Namen wird kein Anspruch auf ein vollständiges Verzeichnis der zur Komponentenanalyse vorliegenden Arbeiten erhoben. Weitere Quellenangaben befinden sich in den Abschnitten 3 und 4, Seite 84 ff.

Neben diesen inhaltlichen Fragen bleibt zunächst auch der wichtige Gesichtspunkt unberührt, ob die n Zahlenreihen im strengen Sinne Zufallsvariable ⁵⁾ mit einer fest zugeordneten Wahrscheinlichkeit für jedes im möglichen Wertebereich liegende X_{ij} , das heißt mit einem gegebenen Verteilungsgesetz sind. Im Zusammenhang damit ist es vorerst ebenfalls unerheblich, ob etwa die N Elemente zum Zwecke der "Messung" durch die Variablen als Stichprobe aus einer größeren Gesamtheit von Elementen ausgewählt worden sind. Als eigentlich statistische Fragestellung wird dagegen angenommen, daß die Variablen im Regelfalle untereinander korreliert sind.

Nun hat die Komponentenganalyse die Aufgabe, den Komplex der (nN) Beobachtungswerte durch Vereinfachung seiner Struktur einer Bearbeitung und Erklärung zugänglicher zu machen. Zu diesem Zweck wollen wir die vorgegebenen Variablen X_i durch eine Anzahl p von neuen Variablen Z_k ($k=1,2,\dots,p$) so ersetzen, daß ein System von Beziehungen der Form

$$(1) \quad X_i = f_i(Z_1, Z_2, \dots, Z_p) \quad (i=1,2,\dots,n)$$

entsteht.

111 Normierung der Variablen

Die allgemeine Gültigkeit der folgenden Ausführungen bleibt unberührt, wenn man die gegebenen Variablen so normiert, daß ihre arithmetischen Mittel 0 und ihre Varianzen 1 werden. Gleiches gilt, wenn dies bei den noch unbekannten Variablen Z_k vorausgesetzt wird. Wir bestimmen daher

$$(2) \quad \bar{X}_i = \frac{1}{N} \sum_{j=1}^N X_{ij} \quad (i=1,2,\dots,n)$$

5) Vgl. ANDERSON, O. (1957³) - Probleme der statistischen Methodenlehre in den Sozialwissenschaften. Würzburg: Physica-Verlag, Seite 90-91.

und

$$(3) \quad \text{Var } X_i = \frac{1}{N} \sum_{j=1}^N (X_{ij} - \bar{X}_i)^2.$$

Dann erhalten wir mit der Transformation

$$(4) \quad x_{ij} = \frac{1}{\sqrt{\text{Var } X_i}} (X_{ij} - \bar{X}_i)$$

die einzelnen Werte der normierten Variablen x_i , die demzufolge durch

$$(5) \quad \bar{x}_i = 0 ; \text{Var } x_i = 1$$

gekennzeichnet sind. Denkt man sich den gleichen Vorgang für die Variablen Z_k , so ist

$$(6) \quad \bar{z}_k = 0 ; \text{Var } z_k = 1 \quad (k=1,2,\dots,p).$$

Wir wollen an dieser Stelle die Auswirkung der Normierung auf den einfachen, linearen Korrelationskoeffizienten zwischen zwei Variablen X_i und X_h betrachten. Bekanntlich ist

$$(7) \quad r_{ih} = \frac{\sum_{j=1}^N (X_{ij} - \bar{X}_i)(X_{hj} - \bar{X}_h)}{N \sqrt{\text{Var } X_i} \sqrt{\text{Var } X_h}} \\ (i,h=1,2,\dots,n).$$

Dem entspricht wegen (5) und unter Berücksichtigung des Umstands, daß der Korrelationskoeffizient gegenüber einer Transformation nach (4) invariant ist, der Ausdruck ⁶⁾

$$(8) \quad r_{ih} = \frac{1}{N} \sum_{j=1}^N x_{ij} x_{hj}.$$

Demgegenüber ist

6) Man beachte, daß die Normierung außerdem die Beziehungen $r_{ih} = \text{Kov}(x_i; x_h)$, $r_{km} = \text{Kov}(z_k; z_m)$ sowie $r_{ii} = \text{Var } x_i = r_{kk} = \text{Var } z_k (= 1)$ zur Folge hat.

$$(9) \quad r_{km} = \frac{1}{N} \sum_{j=1}^N z_{kj} z_{mj} \quad (k, m = 1, 2, \dots, p).$$

112 Die Komponenten und ihre Beziehungen zu den Ausgangsvariablen

Von den gesuchten Variablen z_k wird gefordert, daß sie im Gegensatz zu den x_i untereinander nicht korreliert sein sollen. Daher ist (9) zu ergänzen auf ⁷⁾

$$(10) \quad r_{km} = \delta_{km} = \begin{cases} 1 & \text{für } k = m \\ 0 & \text{für } k \neq m. \end{cases}$$

Die angestrebte Vereinfachung der Ausgangsgesamtheit kann grundsätzlich in diesem Fehlen von Korrelationen, also einer statistischen Unabhängigkeit der z_k gesehen werden, wobei darin das Kriterium einer "fundamentalen" Bedeutung der gesuchten Variablen besteht ⁸⁾. Ebenso wesentlich erscheint in diesem Zusammenhang die Absicht, die Anzahl p der zu bestimmenden Variablen möglichst kleiner als n zu halten ($p < n$). Es ist ersichtlich, daß dies in der Tat eine Vereinfachung der Ausgangsgesamtheit bewirken kann. Wir folgen der von HOTELLING eingeführten Benennung und bezeichnen die Variablen z_k als die Komponenten (components) des Komplexes der (nN) Beobachtungswerte.

Eine weitere Annahme geht dahin, daß zwischen den Variablen x_i und den Komponenten z_k lineare Beziehungen bestehen sollen. Die Voraussetzung funktionaler Zusammenhänge höheren

7) Ist für $(k, m=1, 2, \dots, p)$ ein Ausdruck gleich 0 für alle $(k \neq m)$ und gleich 1 für alle $(k=m)$, so kennzeichnet man dies durch das KRONECKER-Symbol δ_{km} . Davon wollen wir auch in Zukunft Gebrauch machen.

8) Die eigentliche Bedeutung dieser Frage wird später ausführlicher behandelt werden. Vorläufig begnügen wir uns mit der Aussage, daß Variable gesucht werden, die - um die gebräuchlichsten Ausdrücke der Literatur zu benutzen - in einem noch nicht näher definierten Sinne "more fundamental" oder "more convenient" sind.

präzise mathematische Formulierung zu geben ¹⁰⁾. Wir beginnen damit, Übertragungen in die Vektoralgebra vorzunehmen.

Die Variablen x_i und die Komponenten z_k sollen von nun an als N-dimensionale Vektoren aufgefaßt werden. Dies führt zu der Schreibweise

$$\mathbf{r}_i = \{x_{i1} \ x_{i2} \ \dots \ x_{iN}\}$$

und

$$\mathbf{z}_k = \{z_{k1} \ z_{k2} \ \dots \ z_{kN}\}.$$

Das System (11) geht in eine Reihe von Vektorgleichungen der Form

$$\begin{aligned} \mathbf{r}_1 &= a_{11}\mathbf{z}_1 + a_{12}\mathbf{z}_2 + \dots + a_{1p}\mathbf{z}_p \\ \mathbf{r}_2 &= a_{21}\mathbf{z}_1 + a_{22}\mathbf{z}_2 + \dots + a_{2p}\mathbf{z}_p \\ \mathbf{r}_n &= a_{n1}\mathbf{z}_1 + a_{n2}\mathbf{z}_2 + \dots + a_{np}\mathbf{z}_p \end{aligned} \quad (12)$$

oder

$$\mathbf{r}_i = \sum_{k=1}^p a_{ik}\mathbf{z}_k \quad (i=1,2,\dots,n) \quad (12')$$

über.

Dies ist vorerst nur ein formaler Wechsel der Symbole, denn in der Tat können die Vektoren $\mathbf{r}_i$ und $\mathbf{z}_k$ nach der Regel für die Addition von Vektoren sowie der der Multiplikation eines Vektors mit einer Zahl a_{ik} nach (12) zueinander in Beziehung gesetzt werden. Wir wollen nun klären, welche Bedeutung den Eigenschaften der normierten Variablen in der Vektoralgebra zukommt.

10) Eine kurze Zusammenfassung der linearen Algebra, insbesondere der Vektoralgebra, bietet das Kapitel 1 bei CORSTEN, L.C.A. (1958) - Vectors, a Tool in Statistical Regression Theory (Mededelingen van de Landbouwhogeschool te Wageningen, 58/1). Wageningen: Veenman & Zonen, Seite 4-22.

Aus (5) und (8) läßt sich der zusammenfassende Ausdruck

$$(13) \quad \frac{1}{N} \sum_{j=1}^N x_{ij} x_{hj} = \begin{cases} \text{Var } x_i = 1 & \text{für } i=h \\ r_{ih} & \text{für } i \neq h \end{cases}$$

bilden. Dies stimmt aber bis auf einen Faktor $\frac{1}{N}$ mit der Definition des inneren Produktes der Vektoren $\mathbf{e}_i$ und $\mathbf{e}_h$ überein, denn es ist

$$(14) \quad (\mathbf{e}_i, \mathbf{e}_h) = \sum_{j=1}^N x_{ij} x_{hj}.$$

Daraus schließen wir, daß

$$(15) \quad (\mathbf{e}_i, \mathbf{e}_h) = \begin{cases} N \text{Var } x_i = N & \text{für } i=h \\ N r_{ih} & \text{für } i \neq h. \end{cases}$$

Wird die geometrische Veranschaulichung herangezogen, so ermöglicht (15) die folgenden Aussagen über die Länge der Vektoren $\mathbf{e}_i$ und den von ihnen eingeschlossenen Winkel α_{ih} . Für den Fall ($i=h$) gelten

$$(16) \quad (\mathbf{e}_i, \mathbf{e}_i) = N \text{Var } x_i = N$$

und als Betrag oder Länge des Vektors $\mathbf{e}_i$

$$(17) \quad |\mathbf{e}_i| = \sqrt{(\mathbf{e}_i, \mathbf{e}_i)} = \sqrt{N}.$$

Die Vektoren $\mathbf{e}_i$ besitzen also übereinstimmend die Länge $\sqrt{N}$. Wenden wir bei ($i \neq h$) den Ausdruck ¹¹⁾

$$(18) \quad (\mathbf{e}_i, \mathbf{e}_h) = |\mathbf{e}_i| |\mathbf{e}_h| \cos \alpha_{ih}$$

an, so ist wegen (15) und (17)

$$(19) \quad r_{ih} = \cos \alpha_{ih};$$

11) FEIGL, G. (Hrsg. ROHRBACH, H.) (1953) - Einführung in die höhere Mathematik, Berlin (Göttingen, Heidelberg): Springer-Verlag, Seite 55.
SPERNER, E. (1957³) - Einführung in die Analytische Geometrie, Teil 1, Göttingen: Vandenhoeck & Ruprecht, Seite 79-81.
Die folgende Gleichung (19) läßt sich in der Ebene ($N=2$) mit Hilfe der Additionstheoreme für Winkelfunktionen leicht beweisen.

man kann daher sagen, daß der Korrelation zwischen zwei Variablen der Kosinus des von ihren Vektoren eingeschlossenen Winkels entspricht.

Für die Komponenten z_k und ihre Vektoren $\mathfrak{z}_k$ sind diese Beziehungen wegen (10) einfacher:

$$(20) \quad (\mathfrak{z}_k, \mathfrak{z}_m) = N \delta_{km} = \begin{cases} N & \text{für } k=m \\ 0 & \text{für } k \neq m. \end{cases}$$

Die $\mathfrak{z}_k$ haben ebenfalls die Länge $\sqrt{N}$; da sie untereinander nicht korreliert sein sollen, bilden sie - wie (18) und (19) sinngemäß zeigen - paarweise einen rechten Winkel, sie sind orthogonal.

121 Vektorraum und euklidischer Raum

Fügen wir zu den n Vektoren $\mathcal{C}_i$ - mit je N Komponenten ¹²⁾ - eine Anzahl N von linear unabhängigen Vektoren $\mathfrak{n}_j$ ($j=1,2,\dots,N$), die ebenfalls N Komponenten besitzen sollen, hinzu, so sind wir berechtigt, die Gesamtheit der $\mathcal{C}_i$ als ein lineares Vektorgebilde des N -dimensionalen Vektorraumes, den wir mit V_N bezeichnen wollen, aufzufassen. Die $\mathfrak{n}_1, \mathfrak{n}_2, \dots, \mathfrak{n}_N$ bilden dann die Basis des V_N , in welchem je $(N+1)$ Vektoren linear abhängig, das heißt als Linearkombination der $\mathfrak{n}_j$ darstellbar sein müssen. Führen wir dies für den Vektor $\mathcal{C}_i$ durch, so ergibt sich

$$(21) \quad \mathcal{C}_i = x_{i1}\mathfrak{n}_1 + x_{i2}\mathfrak{n}_2 + \dots + x_{iN}\mathfrak{n}_N = \sum_{j=1}^N x_{ij}\mathfrak{n}_j.$$

12) Hier ist eine Doppelsinnigkeit nicht zu vermeiden, da unter "Komponenten" jetzt die einzelnen, einem Vektor in geordneter Reihenfolge zugehörenden Zahlen x_{ij} zu verstehen sind. Nach einer weiteren Überlegung werden wir in der Lage sein, den im Gesamtzusammenhang dieser Arbeit anders zu deutenden Begriff der "Komponenten" durch die Bezeichnung "Koordinaten" zu ersetzen.

Da wir

$$\pi_j = \{e_{j1} \quad e_{j2} \quad \dots \quad e_{jN}\} \quad (j=1,2,\dots,N)$$

gesetzt haben, erhalten wir

$$\pi_i = \left\{ \sum_{j=1}^N x_{ij} e_{j1} \quad \sum_{j=1}^N x_{ij} e_{j2} \dots \sum_{j=1}^N x_{ij} e_{jN} \right\} \\ (i=1,2,\dots,n).$$

Unter diesen Umständen kann der bereits definierte Vektor π_i nur erhalten bleiben, wenn wir als Basis des V_N die Einheitsvektoren ¹³⁾

$$\pi_j = \{0 \quad 0 \quad \dots \quad 0 \quad 1 \quad 0 \quad \dots \quad 0\} \quad (j=1,2,\dots,N)$$

mit

$$(22) \quad e_{jl} = \delta_{jl} = \begin{cases} 1 & \text{für } j=l \\ 0 & \text{für } j \neq l \end{cases} \quad (j,l=1,2,\dots,N)$$

wählen. Die lineare Unabhängigkeit der so gedeuteten Vektoren π_j läßt sich leicht zeigen. Wir können abschließend erklären, daß die π_i als Vektoren des V_N zu betrachten sind, wobei ihre Komponenten x_{ij} ($i=1,2,\dots,n; j=1,2,\dots,N$) nach (21) durch die vorgegebene Basis $\pi_1, \pi_2, \dots, \pi_N$ des Raumes festgelegt werden.

Gleichzeitig mit der Einführung der Vektoren π_i wurden deren innere Produkte (14) erklärt. Dadurch ist das Vorhandensein einer Metrik angedeutet; es liegt nahe, die Beschaffenheit derselben zu prüfen und für das vorliegende Problem eine weitere Raumvorstellung einzuführen.

Hierzu müssen wir untersuchen, welche Werte die Norm der Vektoren des V_N annehmen kann. Nach (16) ist

$$(23) \quad \|\pi_i\| = (\pi_i, \pi_i) = N,$$

während wegen (22) sicherlich

13) Einheitsvektoren haben die Länge 1 und stehen paarweise aufeinander senkrecht.

$$(24) \quad (\pi_j, \pi_1) = \delta_{j1} = \begin{cases} 1 & \text{für } j=1, \\ 0 & \text{für } j \neq 1, \end{cases} \quad \text{das heißt}$$

$$(24') \quad \|\pi_j\| = (\pi_j, \pi_j) = 1$$

gilt. In jedem Falle ist also

$$(25) \quad \|\pi_i\| > 0; \quad \|\pi_j\| > 0 \quad (i=1, 2, \dots, n; j=1, 2, \dots, N).$$

Man kann deshalb sagen, daß der V_N eine positiv definite hermitesche Metrik besitzt ¹⁴⁾. Da außerdem auf Grund der Anlage der Problemstellung für den Bereich der Vektorkomponenten nur der Körper der reellen Zahlen in Frage kommt, ist der V_N ein euklidischer Raum.

Bestimmen wir, daß alle Vektoren des V_N von dem gleichen Ursprungspunkt aus abgetragen werden (Orts- oder Radiusvektoren), so läßt sich der Raum - freilich nur für ($N \leq 3$) - anschaulich erklären. Da die Einheitsvektoren π_j senkrecht aufeinander stehen und die Länge 1 haben, bilden sie eine Orthonormalbasis des V_N . Diese entspricht einem System kartesischer Koordinaten im anschaulichen Raum, wobei die Komponenten der Vektoren π_i durch die Linearkombination

$$(21') \quad \pi_i = \sum_{j=1}^N x_{ij} \pi_j \quad (i=1, 2, \dots, n)$$

als die Koordinaten ihrer Endpunkte in Bezug auf die N Achsen erscheinen. Wegen dieser Analogie wollen wir die x_{ij} von nun an als die Koordinaten der Vektoren π_i in Bezug auf die vorgegebene Basis $\pi_1, \pi_2, \dots, \pi_N$ bezeichnen. Das sei für ($N=3$) am Beispiel eines Vektors π_1 in der folgenden Abbildung verdeutlicht.

14) Dabei haben wir es unterlassen, speziell die Vorbedingungen für eine hermitesche Metrik zu prüfen. Diese sind nach den Regeln des Rechnens mit Vektoren erfüllt. (25) gilt außerdem nur für π_i beziehungsweise $\pi_j \neq 0$ (Nullvektor), was in unserem System vorausgesetzt wird.

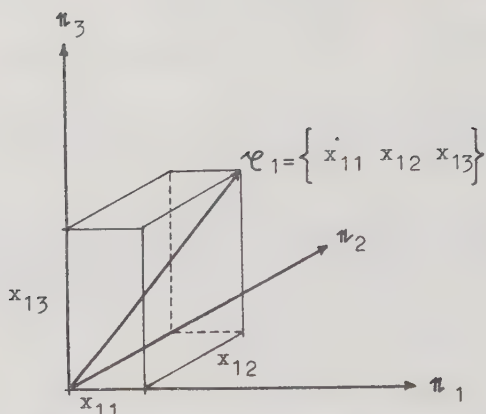


Abb. 1: Die Komponenten eines Ortsvektors als Koordinaten seines Endpunktes im dreidimensionalen Raum.

Auch die p Vektoren $\mathfrak{z}_k$ können als ein lineares Vektorgebilde des V_N angesehen werden. (25) ist wegen (20) erfüllt; (20) zeigt außerdem, daß die $\mathfrak{z}_k$ ein im V_N eingebettetes System von orthogonalen Vektoren darstellen, von denen jeder analog zu (21') mit der Basis n_j durch

$$(26) \quad \mathfrak{z}_k = \sum_{j=1}^N z_{kj} n_j \quad (k=1,2,\dots,p)$$

ausgedrückt werden kann.

Wir beenden diesen Abschnitt mit der Feststellung, daß die Ausgangsvariablen x_i und die Komponenten z_k als Vektoren des N -dimensionalen euklidischen Raumes V_N , für den als Basis die N Einheitsvektoren n_j fest gewählt wurden, gedacht werden können.

122 Die Komponentenanalyse als Problem der linearen Algebra

Die Einführung der vektoriellen Betrachtungsweise schafft die Möglichkeit, die im Abschnitt 110 gestellte Aufgabe der Komponentenanalyse genauer zu formulieren.

Wir gehen von dem vorgegebenen V_N aus und stellen die Frage, welches die größte Zahl von linear unabhängigen Vektoren ist, die sich für die n Vektoren $\mathcal{C}_i$ bestimmen läßt. Diese Zahl, die offenbar höchstens den Wert n annehmen kann, nennen wir p ($p \leq n$). Nun bilden p linear unabhängige Vektoren in einem N -dimensionalen Vektorraum einen p -dimensionalen Teilraum, den wir mit T_p bezeichnen wollen und der in unserem Falle selbst wieder ein euklidischer Raum ist ¹⁵⁾. Dieser Fall kann bei ($n \geq N$) eintreten, insbesondere bei ($n < N$), weil dann wegen ($p \leq n$) ohnehin ein Teilraum gegeben ist. Für die Frage des Teilraumes ist das Verhältnis zwischen N und p maßgebend, und zwar sind die Möglichkeiten ($N > p$) und ($N = p$) denkbar. Ist ($N > p$), so liegt in allen Fällen von ($n \geq N$) ein p -dimensionaler Teilraum vor. Bei ($N = p$) fallen die Räume V_N und T_p zusammen, der V_N ist der durch die Zahl der linear unabhängigen unter den Vektoren $\mathcal{C}_i$ bestimmte Raum selbst ¹⁶⁾. Unsere weiteren Ausführungen werden durch diese Unterscheidungen nicht berührt, da sie sowohl für ($N > p$) als auch für den Sonderfall ($N = p$) gelten.

Als Basis für den T_p kommt zunächst die p -gliedrige Basis der unter den $\mathcal{C}_i$ auffindbaren linear unabhängigen Vektoren mit der Linearkombination ¹⁷⁾

-
- 15) Das ergibt sich daraus, daß (25) auch für irgendwelche p Vektoren unter den $\mathcal{C}_i$ gelten muß.
- 16) Der Fall ($N < p$) scheidet aus, da dann der V_N ein Teilraum des T_p wäre. Dies widerspricht der Voraussetzung, daß alle hier betrachteten Vektoren genau N Koordinaten haben, wodurch die Dimension der Räume höchstens gleich N sein kann. Ferner ist zu beachten, daß bei ($N = p$) nur noch ($n \geq N$) gilt, weil die Ungleichungen ($n < N$) und ($p \leq n$) nicht mit ($N = p$) verträglich sind.
- 17) Da wir uns im N -dimensionalen Raum bewegen, sehen wir von einer Orthonormalbasis $u_1, u_2, \dots, u_p$ mit den p -dimensionalen Einheitsvektoren ab. Für (27), Seite 14, müssen wir statt i den Index h verwenden, weil wir nicht festgelegt haben, welche p Vektoren unter den $\mathcal{C}_i$ linear unabhängig sind. Vgl. formal (7) auf Seite 4.

$$(27) \quad \mathcal{C}_h = \sum_{k=1}^p b_{hk} \mathcal{C}_k \quad (h=1,2,\dots,n)$$

in Frage.

Eine Basis dieser Art genügt jedoch nicht unseren Zwecken. Dagegen geht aus unserer Problemstellung hervor, daß wir jetzt auf die Vektoren $\mathcal{J}_k$ ($k=1,2,\dots,p$), die die gesuchten z_k repräsentieren, zurückgreifen müssen. Für diese Vektoren ist bereits gezeigt worden, daß sie ein p -dimensionales Vektorgebilde darstellen, das wegen seiner Orthogonalität (Fehlen von Korrelationen) und der Vektorlänge $\sqrt{N}$ den Anforderungen der Komponentenanalyse entspricht. Wählen wir die $\mathcal{J}_k$ als Basis, so sind im T_p je $(p+1)$ Vektoren linear abhängig und als Linearkombination der $\mathcal{J}_1, \mathcal{J}_2, \dots, \mathcal{J}_p$ in der Form ¹⁸⁾

$$(12'') \quad \mathcal{C}_i = \sum_{k=1}^p a_{ik} \mathcal{J}_k \quad (i=1,2,\dots,n)$$

darstellbar. Das entspricht aber genau unserer durch (11) oder (11') ausgedrückten Aufgabenstellung.

Damit ist es möglich, das Problem der Komponentenanalyse im Rahmen der linearen Algebra wie folgt zu charakterisieren. Vorgegeben ist der N -dimensionale euklidische Raum V_N mit den Vektoren $\mathcal{C}_i$ ($i=1,2,\dots,n$), deren Koordinaten durch die fest gewählte Orthonormalbasis $u_1, u_2, \dots, u_N$ nach (21') bestimmt werden. Aufgabe ist, für den durch die Maximalzahl p ($p \leq n$) linear unabhängiger Vektoren unter den $\mathcal{C}_i$ aufgespannten Teilraum T_p des V_N eine Orthogonalbasis $\mathcal{J}_1, \mathcal{J}_2, \dots, \mathcal{J}_p$ zu finden und die Koordinaten der $\mathcal{C}_i$ so zu transformieren, daß sie durch (12'') festgelegt sind ¹⁹⁾.

18) Vgl. (12') auf Seite 7.

19) Für den Fall ($N=p$) verändert sich die Aufgabenstellung lediglich dahingehend, daß in einem Raum gleicher Dimension ein Wechsel von der Orthonormalbasis u_j zur Orthogonalbasis $\mathcal{J}_k$ vorzunehmen ist ($j,k=1,2,\dots,N=p$).

Setzt man ($N=3$) und ($p=2$) und wird wieder angenommen, daß alle Vektoren denselben Bezugspunkt haben, so ist die geschilderte Aufgabenstellung in der Geometrie des Raumes deutbar. Wir haben dann ein zweidimensionales Vektorgebilde in einem Raum mit kartesischem Achsensystem (x_1, x_2, x_3), und zwar n Vektoren $\mathcal{C}_i$, die in einer Ebene liegen und deren Koordinaten den Projektionen ihrer Endpunkte auf die drei Achsen entsprechen. Das Problem besteht darin, im Raum die Koordinaten eines rechtwinkligen Achsensystems ($\mathcal{Z}_1, \mathcal{Z}_2$) für die p -Ebene zu bestimmen und die $\mathcal{C}_i$ nach (12'') in dieses zu überführen. Damit können die $\mathcal{C}_i$ durch die orthogonalen $\mathcal{Z}_k$ ersetzt werden, allerdings unter durchgehender Dehnung ihrer Projektionen auf das $\sqrt{3}$ -fache. Das liegt daran, daß auf den Achsen des dreidimensionalen Systems die Meßpunkte die Abstände 1 haben, während sie auf den Achsen der Ebene um jeweils $\sqrt{N} = \sqrt{3}$ voneinander entfernt sind.

13 KOMPONENTENANALYSE UND LINEARE ALGEBRA: MATRIZEN

130 Definition der vorkommenden Matrizen

Die bisher durchgeführte Ableitung der Komponentenanalyse nahm ihren Ausgang mit den Beziehungen (11) beziehungsweise (12). Diese stellen formal ein inhomogenes System von n linearen Gleichungen in p Unbekannten dar ²⁰⁾. Es ist wünschenswert, dieses System mit Hilfe der Konventionen des Matrizenkalküls auf eine Form zu bringen, die seine Darstellung auf drei Symbole beschränkt und die weiteren Operationen erleichtert.

20) Diese Stelle ist geeignet, vorschauend auf die besondere Problematik der Komponentenanalyse hinzuweisen: Im System (11) sind nicht nur die z_k , sondern auch die Koeffizienten a_{ik} unbekannt, da nur die Variablen x_i vorgegeben wurden. Offensichtlich muß die Analyse durch die Wahl dazu geeigneter Bedingungen ihren Anfang in der Bestimmung des Koeffizientenschemas a_{ik} finden.

Gehen wir von (12) aus, so erreichen wir das Ziel bequemer Handhabung durch das Übereinkommen, die Koordinaten der $\mathcal{C}_i$ und $\mathcal{Z}_k$ sowie die Koeffizienten a_{ik} in der folgenden Weise als rechteckige Matrizen zu schreiben:

$$\mathcal{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nN} \end{bmatrix}, \quad \mathcal{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1N} \\ z_{21} & z_{22} & \cdots & z_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ z_{p1} & z_{p2} & \cdots & z_{pN} \end{bmatrix}$$

und

$$\mathcal{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{bmatrix}.$$

Hierbei sind $\mathcal{X}$ und $\mathcal{Z}$ (n, N) - beziehungsweise (p, N) -reihige Matrizen, in denen die Vektoren $\mathcal{C}_i$ und $\mathcal{Z}_k$ jeweils als Zeilen angeordnet sind. Da $\mathcal{A}$ als (n, p) -Matrix mit $\mathcal{Z}$ verkettet, also multiplizierbar ist, kann (12) nach den Regeln des Rechnens mit Matrizen durch die Gleichung

$$(28) \quad \mathcal{X} = \mathcal{A} \mathcal{Z}$$

ersetzt werden.

Es wird später erforderlich sein, die Matrizen in ihre Zeilen oder Spalten zu zerlegen und diese wiederum als Matrizen aufzufassen ²¹⁾. Bei $\mathcal{X}$ und $\mathcal{Z}$ führt dies zur Zerlegung

21) Dabei machen wir von der Möglichkeit Gebrauch, Matrizen als aus einer Reihe von Teilmatrizen (Zeilen- und Spaltenvektoren, aber auch anderen Teilanordnungen und schließlich den einzelnen Elementen selbst) zusammengesetzt zu betrachten. Eine so aufgebaute Matrix heißt Übermatrix. Siehe hierzu THRALL, R.M.; TORNHEIM, L. (1957) - Vector Spaces and Matrices. New York: J. Wiley & Sons, Seite 64-66 sowie GANTMACHER, F.R. (1958) - Matrizenrechnung. Teil I: Allgemeine Theorie. Berlin: VEB Deutscher Verlag der Wissenschaften, Seite 38-46.

in je n beziehungsweise p $(1,N)$ -reihige Matrizen (Zeilen-Vektoren), die den Vektoren $\mathcal{C}_i$ und $\mathcal{Z}_k$ entsprechen:

$$\mathcal{H} = \begin{bmatrix} \mathcal{C}_1 \\ \mathcal{C}_2 \\ \vdots \\ \mathcal{C}_n \end{bmatrix} \quad \text{sowie} \quad \mathcal{Z} = \begin{bmatrix} \mathcal{Z}_1 \\ \mathcal{Z}_2 \\ \vdots \\ \mathcal{Z}_p \end{bmatrix}.$$

Für $\mathcal{A}$ wollen wir eine Zerlegung in Zeilen- und Spaltenvektoren vorsehen und die Darstellung

$$\mathcal{A} = \begin{bmatrix} \mathbf{a}_1^* \\ \mathbf{a}_2^* \\ \vdots \\ \mathbf{a}_n^* \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_p \end{bmatrix}$$

vereinbaren.

Aus (28) lassen sich weitere Matrizengleichungen ableiten. Durch gegenseitiges Vertauschen der Zeilen und Spalten von $\mathcal{H}$ entsteht die zu $\mathcal{H}$ transponierte (N,n) -reihige Matrix

$$\mathcal{H}' = \begin{bmatrix} x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1N} & x_{2N} & \dots & x_{nN} \end{bmatrix}.$$

Linksmultiplikation mit $\mathcal{H}$ ergibt die (n,n) -reihige Produktmatrix

$$(29) \quad \mathcal{H} \mathcal{H}' = \begin{bmatrix} \sum_{j=1}^N x_{1j}^2 & \sum_{j=1}^N x_{1j} x_{2j} & \dots & \sum_{j=1}^N x_{1j} x_{nj} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{j=1}^N x_{nj} x_{1j} & \sum_{j=1}^N x_{nj} x_{2j} & \dots & \sum_{j=1}^N x_{nj}^2 \end{bmatrix},$$

eine symmetrische Matrix mit den Elementen 22)

$$(14') \quad (c_i, c_h) = c_i c'_h = \sum_{j=1}^N x_{ij} x_{hj} \quad (i, h=1, 2, \dots, n).$$

Multiplizieren wir $\mathcal{X}\mathcal{X}'$ mit dem Faktor $\frac{1}{N}$, so erhalten wir wegen (15) die quadratische Matrix

$$(30) \quad \frac{1}{N} \mathcal{X}\mathcal{X}' = \begin{bmatrix} 1 & r_{12} & \dots & r_{1n} \\ r_{21} & 1 & \dots & r_{2n} \\ \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & \dots & 1 \end{bmatrix} = \mathcal{R}$$

mit der Hauptdiagonalen

$$(31) \quad 1 = \text{Var} c_i = r_{ii} = \frac{1}{N} c_i c'_i.$$

Da diese Matrix die Korrelationskoeffizienten zwischen den Ausgangsvariablen x_i enthält, soll sie Korrelationsmatrix $\mathcal{R}$ genannt werden 23).

Analog zu (29) erhält man für das Produkt $\mathcal{Z}\mathcal{Z}'$ eine symmetrische (p,p)-Matrix mit den Elementen

22) In der Schreibweise $c_i = \{x_{i1} \ x_{i2} \ \dots \ x_{iN}\}$ war offen geblieben, ob es sich um einen Zeilen- oder Spaltenvektor handelt. Nun haben wir c_i als Zeilenvektor von $\mathcal{X}$ definiert, infolgedessen ist c'_i als zu c_i transponierter Vektor ein Spaltenvektor. Im Matrizenkalkül ist skalare (innere) Multiplikation nur zeilen- und spaltenweise sinnvoll erklärt. An die Stelle des Produkts (c_i, c_h) tritt daher als Ausdruck für die Multiplikation einer $(1, N)$ - mit einer $(N, 1)$ -Matrix die $(1, 1)$ -reihige Matrix $c_i c'_h$. Da $\mathcal{X}\mathcal{X}'$ symmetrisch ist, gilt im übrigen $c_i c'_h = c'_h c_i$.

23) Als Folge der Normierung enthält sie gleichzeitig in der Hauptdiagonalen die Varianzen, sonst die Kovarianzen der Variablen (vgl. Fußnote 6, Seite 4). $\mathcal{R}$ kann deshalb auch als Varianz-Kovarianz-Matrix, Streuungs- oder Dispersionsmatrix bezeichnet werden. Vgl. DUBOIS, P.H. (1957) - Multivariate Correlational Analysis. New York: Harper & Brothers, Seite 74-79. KENDALL-BUCKLAND (1957), a.a.O., Seite 69.

$$(20') \quad (z_k, z_m) = z_k z'_m = \sum_{j=1}^N z_{kj} z_{mj} \quad (k, m=1, 2, \dots, p)$$

und deshalb wegen der durch (10) festgehaltenen Voraussetzung unkorrelierter Komponenten mit

$$(32) \quad \frac{1}{N} z z' = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} = \mathcal{E}$$

die p-reihige Einheitsmatrix als "Korrelationsmatrix" der z_k .

131 Dimension und Rang des Komponentenproblems

Da die Bedeutung der Matrizenschreibweise ausschließlich in der Vereinfachung komplizierter linearer Beziehungen besteht und diese, wie in den Abschnitten 120 und 121 gezeigt wurde, in die mehrdimensionale Geometrie übertragen werden können, muß eine eindeutige Zuordnung zwischen Matrizen und Vektorräumen bestehen. Dabei entspricht bekanntlich der Dimension eines Vektorraumes - also der Maximalzahl linear unabhängiger Vektoren - der Begriff des Ranges der aus den Vektorkoordinaten gebildeten Matrix. Demnach ist der Rang einer Matrix gleich der größten Zahl der unter den Reihen auffindbaren linear unabhängigen Zeilen oder Spalten ²⁴⁾.

Eine Betrachtung der (n,N)-Matrix $\mathcal{Z}$ zeigt, daß rein formal ihr Rang höchstens gleich der kleineren der beiden Zahlen n oder N sein kann ²⁵⁾. Dies entspricht durchaus der

24) Zeilenrang und Spaltenrang einer Matrix stimmen stets überein, der Rang ist gegenüber dem Vertauschen von Zeilen und Spalten invariant. Siehe ZURMÜHL, R. (1958²) - Matrizen. Berlin (Göttingen, Heidelberg): Springer-Verlag, Seite 86-90.

25) ZURMÜHL, R. (1958), a.a.O., Seite 90, Satz 2.

Vorstellung, die wir von der Dimension p des von den $\mathcal{C}_i$ aufgespannten Vektorgebildes hatten; überhaupt sind alle Ausführungen ²⁶⁾ über die Dimension des V_N beziehungsweise T_p uneingeschränkt auf den Rang der Matrix $\mathcal{R}$ übertragbar, wonach $\mathcal{R}$ als Rang die Zahl p mit $(p \leq n)$ und $(p \leq N)$ zuzuordnen ist.

Aus einer allgemeinen Regel des Matrizenkalküls läßt sich folgern ²⁷⁾, daß der Rang der Produktmatrix bei Multiplikation einer Matrix mit ihrer Transponierten, wie im Falle (30) und (32) gegeben, unverändert als Rang der Faktoren erhalten bleibt. Daraus entnehmen wir, daß auch für $\mathcal{R}$ der Rang von $\mathcal{R}$ gelten muß. Der Rang p der Korrelationsmatrix ist also Ausdruck der kleinsten Dimension, auf die unsere n -dimensionale Ausgangsstruktur reduziert werden kann. Sind aber sämtliche Zeilen beziehungsweise Spalten von $\mathcal{R}$ linear unabhängig (Rang $p=n$), so kann die Dimension nicht verringert werden; die Zahl der Komponenten- z_k ist dann ebenso groß wie die der Ausgangsvariablen x_i , und der Vorteil der Komponentenanalyse beschränkt sich vorläufig auf die Möglichkeit, n korrelierte Variable durch $(p=n)$ unkorrelierte Komponenten ersetzen zu können. Für (32) folgt übrigens der Rang p schon daraus, daß die p Einheitsvektoren von $\mathcal{E}$, wie wir wissen, linear unabhängig sind.

14 DETERMINATION DES PROBLEMS DURCH HAUPTKOMPONENTEN

140 Die Einführung von Hauptkomponenten

In der bisher erarbeiteten Gestalt ist das Komponentenproblem erst dann eindeutig lösbar, wenn an die gesuchte Orthogonalbasis zusätzliche Forderungen gestellt werden. Man

26) Vgl. Seite 13.

27) Siehe zum Beispiel ZURMÜHL, R. (1958), a.a.O., Seite 92. Vgl. auch Fußnote 24, Seite 19.

sieht, daß die Gleichung (28) mit unbekannten Matrizen α und $\mathfrak{Z}$ von unendlich vielen Basissystemen $\mathfrak{z}_1, \mathfrak{z}_2, \dots, \mathfrak{z}_p$ mit jeweils wechselnder Koeffizientenmatrix α erfüllt werden kann. Keines dieser Systeme kann den Vorrang vor einem anderen beanspruchen. Das ist gleichbedeutend mit der Feststellung, daß ein System von p Achsen gegenüber einem gegebenen Vektorgebilde in unendlich viele Lagen gebracht werden kann, wobei sich jeweils die Koordinaten der Vektoren verändern. Unser Problem ist also nicht determiniert. Wir müssen einen Auswahlgrundsatz finden, mit dessen Hilfe eines der möglichen Basissysteme bevorzugt werden kann; so dann sind für dieses die Koeffizienten a_{ik} zu bestimmen. Dazu nehmen wir die Gelegenheit, eine statistische Fragestellung einzuführen.

Als Kriterium für die eindeutige Lösung wählen wir die Forderung, daß die Komponente $\mathfrak{z}_1$ den größtmöglichen Anteil an den Varianzen der $\mathfrak{r}_1, \mathfrak{r}_2, \dots, \mathfrak{r}_n$ auf sich vereinigen soll. Für die folgenden Komponenten $\mathfrak{z}_2, \mathfrak{z}_3, \dots, \mathfrak{z}_p$ hat dementsprechend zu gelten, daß ihre Anteile an den jeweiligen Residualvarianzen der $\mathfrak{r}_i$ zu maximieren sind. Diejenigen unter allen möglichen $\mathfrak{z}_k$, die dieser Bedingung nachkommen, nennen wir die Hauptkomponenten (principal components) ²⁸⁾ des Systems der n Ausgangsvariablen.

Zu dieser Zielsetzung ist folgendes zu bemerken. Wir haben in 131 festgestellt, daß das n -dimensionale Komponentenproblem nur dann auf eine kleinere Dimension p reduziert werden kann, wenn die Matrix $\mathcal{R}$ den Rang p ($p < n$) hat. Nun zeigt die praktische Erfahrung, daß empirische Korrelationsmatrizen in aller Regel nichtsingulär mit dem Rangabfall ²⁹⁾

28) Diese Bezeichnung findet sich zuerst bei HOTELLING, H. (1933), a.a.O., Seite 421.

29) Eine quadratische Matrix heißt singulär, wenn ihre Zeilen (und Spalten) linear abhängig sind. Im Falle linearer Unabhängigkeit der Reihen heißt sie nichtsingulär oder regulär. Als Rangabfall oder Defekt d bezeichnet man die Differenz zwischen der Reihenzahl und dem Rang einer quadratischen Matrix.
Vgl. ZURMÜHL, R. (1958), a.a.O., Seite 14-15.

($d=0$) sind. Wenn also der Fall ($p=n$) nur das Ersetzen von n Ausgangsvariablen durch ebenfalls n Komponenten gestattet - die Zahl der erklärenden Vektoren somit nicht verringert werden kann -, so bietet die Wahl von Hauptkomponenten die Aussicht, daß die aufeinanderfolgende Maximierung der Varianzanteile den Variationsbereich der $\mathcal{C}_i$ nahezu vollständig ausschöpft, ehe die volle Zahl n der Vektoren $\mathcal{Z}_k$ bestimmt worden ist. Inwieweit die nur zum Teil vollständige Erfassung der Gesamtvariation als hinreichend anzusehen ist, muß der jeweiligen Fragestellung überlassen bleiben. Es kann jedoch gesagt werden, daß die Einführung von Hauptkomponenten neben der von ihnen bewirkten Eindeutigkeit der Lösung eine Möglichkeit schafft, auch bei nichtsingulärer Matrix $\mathcal{R}$ die Zahl der effektiv benötigten Vektoren $\mathcal{Z}_k$ klein zu halten. Eine zusätzliche Vereinfachung ergibt sich dadurch natürlich auch für den Fall ($p < n$).

Die geometrische Bedeutung der Hauptkomponenten bedarf einer genaueren Untersuchung; diese kann jedoch erst später - im Abschnitt 161 - vorgenommen werden. Aus den bisherigen Ausführungen geht hervor, daß die Hauptkomponenten nacheinander zu behandeln sind; in der Tat werden wir im folgenden nicht zuerst die Hauptkomponenten - wie am Beginn dieses Abschnitts angegeben -, sondern die Koordinaten $a_{11}, a_{21}, \dots, a_{n1}$ der Vektoren $\mathcal{C}_i$ ($i=1,2,\dots,n$) in Bezug auf einen Basisvektor $\mathcal{Z}_1$, dann die Koordinaten der $\mathcal{C}_i$ in Bezug auf einen zu $\mathcal{Z}_1$ orthogonalen Vektor $\mathcal{Z}_2$, und so fort, ermitteln. Nach den im Abschnitt 130 getroffenen Vereinbarungen entspricht dies der aufeinanderfolgenden Bestimmung der Spalten der Matrix $\mathcal{A}$, nämlich der n -dimensionalen Vektoren $\mathcal{a}_1, \mathcal{a}_2, \dots, \mathcal{a}_p$. Ist schließlich $\mathcal{A}$ bestimmt, so können aus (28) die in unserem Sinne optimal ausgestalteten Hauptkomponenten $\mathcal{Z}_k$ errechnet und an die Stelle der $\mathcal{C}_i$ gesetzt werden.

141 Hauptkomponenten und Varianzen der Ausgangsvariablen

Wenn nun die Hauptkomponenten an den Varianzen aller $\mathcal{C}_i$ einen jeweils in Bezug auf die verbleibende Gesamtvarianz maximalen Anteil haben sollen und dies durch eine entsprechende Festlegung der Koeffizienten a_{ik} zu bewerkstelligen ist, gilt es jetzt, die Varianzen der $\mathcal{C}_i$ durch die Zeilen und Spalten der Matrix $\mathcal{A}$ auszudrücken.

Wir kehren zurück zu

$$(28') \quad \mathcal{X} = \mathcal{A} \mathcal{Z}$$

und

$$(30') \quad \mathcal{R} = \frac{1}{N} \mathcal{X} \mathcal{X}'.$$

Durch Einsetzen erhält man

$$(33) \quad \mathcal{R} = \frac{1}{N} (\mathcal{A} \mathcal{Z}) (\mathcal{A} \mathcal{Z})', \quad \text{wobei } (30)$$

$$(34) \quad (\mathcal{A} \mathcal{Z})' = \mathcal{Z}' \mathcal{A}'.$$

Also ist

$$(35) \quad \mathcal{R} = \frac{1}{N} \mathcal{A} \mathcal{Z} \mathcal{Z}' \mathcal{A}' = \mathcal{A} \frac{1}{N} \mathcal{Z} \mathcal{Z}' \mathcal{A}'$$

mit

$$(32') \quad \frac{1}{N} \mathcal{Z} \mathcal{Z}' = \mathcal{I}.$$

Daraus folgt

$$(36) \quad \mathcal{R} = \mathcal{A} \mathcal{A}' = \left[\begin{array}{cccc} \sum_{k=1}^p a_{1k}^2 & \sum_{k=1}^p a_{1k} a_{2k} & \cdots & \sum_{k=1}^p a_{1k} a_{nk} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{k=1}^p a_{nk} a_{1k} & \sum_{k=1}^p a_{nk} a_{2k} & \cdots & \sum_{k=1}^p a_{nk}^2 \end{array} \right].$$

30) Die Transponierte einer Produktmatrix ist gleich dem Produkt der transponierten Faktoren in umgekehrter Reihenfolge.

Ein Vergleich der symmetrischen (n,n) -Matrix aa' mit $\mathcal{R}$ in der Form (30) liefert wegen (36) die Beziehungen

$$(37) \quad r_{ih} = \begin{cases} \text{Var } \mathcal{P}_i = \sum_{k=1}^p a_{ik}^2 = 1 & \text{für } i=h \\ \text{Kov } (\mathcal{P}_i, \mathcal{P}_h) = \sum_{k=1}^p a_{ik} a_{hk} & \text{für } i \neq h \end{cases}$$

($i, h=1, 2, \dots, n$).

Damit können die Varianzen der $\mathcal{P}_i$ durch das System

$$(38) \quad \begin{aligned} \text{Var } \mathcal{P}_1 &= a_{11}^2 + a_{12}^2 + \dots + a_{1p}^2 \\ \text{Var } \mathcal{P}_2 &= a_{21}^2 + a_{22}^2 + \dots + a_{2p}^2 \\ &\vdots \\ \text{Var } \mathcal{P}_n &= a_{n1}^2 + a_{n2}^2 + \dots + a_{np}^2 \end{aligned}$$

ausgedrückt werden.

In jeder der Quadratsummen auf den rechten Seiten von (38) befindet sich genau einmal ein Glied, das über den Koeffizienten a_{ik} nach (11) zur Hauptkomponenten $\mathcal{Z}_k$ ($k=1, 2, \dots, p$) gehört. Es folgt daraus, daß der Anteil von $\mathcal{Z}_k$ an allen Varianzen der $\mathcal{P}_i$ - den wir mit S_k bezeichnen wollen - durch spaltenweises Lesen von (38) mit

$$(39) \quad S_k = a_{1k}^2 + a_{2k}^2 + \dots + a_{nk}^2 = \sum_{i=1}^n a_{ik}^2$$

($k=1, 2, \dots, p$)

gegeben ist. Bei Benutzung der zerlegten ³¹⁾ Matrix α ist

$$(37') \quad \text{Var } \mathcal{P}_i = \alpha_{i1}^* \alpha_{i1}^* \quad (i=1, 2, \dots, n),$$

aber

$$(39') \quad S_k = \alpha_k' \alpha_k \quad (k=1, 2, \dots, p).$$

31) Vgl. Seite 17.

Beziehen wir uns auf Hauptkomponenten und behandeln wir den Vektor $\mathbf{z}_1$, so ist

$$(40) \quad S_1 = \mathbf{u}_1' \mathbf{u}_1 = \text{Max} !$$

die zu maximierende Größe.

142 Vorbereitung der Extremwertaufgabe: Nebenbedingungen

Vor der Differentiation von S_1 ist zu prüfen, ob bei der Lösung des Maximumproblems (40) zusätzliche Bedingungen zu beachten sind. Diese könnten darin bestehen, daß die Koordinaten der Vektoren $\mathbf{u}_1^*$ bereits an anderer Stelle unserer Analyse in einer Weise aufgetreten sind, die den Geltungsbereich eines Maximums für S_1 einschränkt.

Daß solche Bedingungen vorliegen, zeigt eine Betrachtung von (36) und (37). Im Anschluß an (37') kann nämlich unter Hinzuziehung der nicht in der Hauptdiagonalen von $\mathbf{a}\mathbf{a}'$ liegenden Werte die Gleichung

$$(37'') \quad r_{ih} = \sum_{k=1}^p a_{ik} a_{hk} = \mathbf{u}_i^* \mathbf{u}_h^* \quad (i, h=1, 2, \dots, n)$$

aufgestellt werden. Das heißt aber, daß nicht nur die Varianzen der $\mathbf{e}_i$, sondern auch die Korrelationen zwischen den $\mathbf{e}_i$ durch die a_{ik} ausgedrückt werden können ³²⁾. Diese sind vorgegeben, deshalb muß (37'') gegenüber der Maximierung von (40) invariant bleiben. Wir hatten durch (19) gesehen, daß die Korrelationen dem Kosinus der Winkel zwischen den $\mathbf{e}_i$ entsprechen. So ist (37'') auch dadurch verständlich, daß das vorgegebene Vektorgebilde der $\mathbf{e}_i$ in der gegenseitigen Stellung der Vektoren zueinander durch einen Wechsel des Basissystems (Koordinatensystems) nicht verändert werden darf. Als Folge der Symmetrie von $\mathbf{R}$ gibt es $\frac{1}{2}n(n+1)$

32) Davon machen wir erst jetzt direkt Gebrauch, denn an sich gilt dies wegen $r_{ii} = \text{Var} \mathbf{e}_i$ auch für die Hauptdiagonale von $\mathbf{a}\mathbf{a}'$. Siehe auch Fußnote 6 auf Seite 4.

verschiedene Werte für (37''). Unser Ziel besteht darin, unter Beachtung dieser zusätzlichen Bedingungen einen Vektor $\mathfrak{z}_1$ so zu bestimmen, daß S_1 ein Maximum wird.

Der Gleichung (37'') kann eine weitere Bedeutung beigelegt werden, die uns später noch beschäftigen wird. Fordern wir nämlich, daß die Komponenten die Korrelationen zwischen den Ausgangsvariablen reproduzieren sollen, so erfüllen genau die von uns definierten Vektoren $\mathfrak{z}_k$ - wie (36) und die Beziehungen (37) zeigen - diese Bedingung. Diese grundsätzlich andere Betrachtungsweise spielt in den Modellen der Faktoranalyse eine gewisse Rolle.

143 Lösung der Extremwertaufgabe

Durch die Beachtung zusätzlicher Bedingungen haben wir unser Problem auf eine Form gebracht, die in der Differentialrechnung als Aufgabe der Bestimmung von Extremwerten für eine Funktion mehrerer Veränderlicher bei Vorhandensein von Nebenbedingungen bekannt ist ³³⁾. Als Lösungsverfahren existiert hierzu die Multiplikatorenmethode, auch Methode von LAGRANGE genannt.

Zur Anwendung dieses Verfahrens ist es erforderlich, aus der Funktion $S_1(a_{11}, a_{21}, \dots, a_{n1})$ unter Hinzunahme der durch die Nebenbedingungen gegebenen Funktionen (37'') einen zusammengesetzten Ausdruck zu bilden. Dabei sind die Nebenbedingungen r_{ih} mit je einem unbestimmten Faktor x_{ih}

33) Vgl. etwa
BIEBERBACH, L. (1944⁵) - Differentialrechnung (Teubners Mathematische Leitfäden, Bd. 4), Leipzig (Berlin): B.G. Teubner, Seite 138-140.
MANGOLDT, H. v. (Bearb. KNOPP, K.) (1948⁹) - Einführung in die höhere Mathematik, Bd. 2, Stuttgart: S. Hirzel Verlag, Seite 398-402.
DUSCHEK, A. (1950) - Vorlesungen über höhere Mathematik, Bd. 2, Wien: Springer-Verlag, Seite 141-144.

(LAGRANGESCHER Multiplikator) zu versehen. Wir schreiben ³⁴⁾

$$(41) \quad 2T_1 = S_1 - \chi_{11}r_{11} - \chi_{12}r_{12} - \dots - \chi_{nn}r_{nn}$$

und erhalten in

$$(42) \quad T_1 = \frac{1}{2} \alpha'_1 \alpha_1 - \frac{1}{2} \sum_{i,h=1}^n \chi_{ih} \alpha_i^* \alpha_h^*$$

eine zusammengesetzte Funktion $T_1(a_{11}, a_{12}, \dots, a_{np}, r_{11}, r_{12}, \dots, r_{nn})$, die so zu behandeln ist, als sei für sie ein Extremwert gesucht. Die dafür notwendige Bedingung geht dahin, die ersten partiellen Ableitungen nach allen Werten a_{ik} zu bilden und ihre Nullstellen zu bestimmen:

$$(43) \quad \frac{\partial T_1}{\partial a_{ik}} = 0 \quad (i=1, 2, \dots, n; k=1, 2, \dots, p).$$

Es ist zweckmäßig, diesen Prozeß in zwei Stufen einzuteilen und zunächst die partiellen Differentialquotienten von T_1 nach den Größen a_{i1} , dann nach allen anderen a_{ik} ($k \neq 1$) zu bestimmen.

Wir ordnen die Differentialquotienten wie folgt nach Spaltenvektoren an

$$\mathcal{J} = [\mathcal{J}_1 \mathcal{J}_2 \dots \mathcal{J}_p] = \begin{bmatrix} \frac{\partial T_1}{\partial a_{11}} & \frac{\partial T_1}{\partial a_{12}} & \dots & \frac{\partial T_1}{\partial a_{1p}} \\ \hline \frac{\partial T_1}{\partial a_{n1}} & \frac{\partial T_1}{\partial a_{n2}} & \dots & \frac{\partial T_1}{\partial a_{np}} \end{bmatrix}$$

und definieren eine (n,n) -reihige sowie wegen $(\chi_{ih} = \chi_{hi})$ symmetrische Matrix mit den LAGRANGESCHEN Multiplikatoren χ_{ih} :

34) Aus Gründen der vereinfachten Darstellung werden durch (41) n^2 Nebenbedingungen gesetzt. Wegen $r_{ih} = r_{hi}$ gilt $\chi_{ih} = \chi_{hi}$.

$$\mathcal{K} = \begin{bmatrix} \kappa_{11} & \kappa_{12} & \dots & \kappa_{1n} \\ \kappa_{21} & \kappa_{22} & \dots & \kappa_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \kappa_{n1} & \kappa_{n2} & \dots & \kappa_{nn} \end{bmatrix}.$$

Diese Vereinbarungen führen nach Durchführung der Differentiation zu den Gleichungen ³⁵⁾

$$(44) \quad \mathcal{J}_1 = \frac{\partial T_1}{\partial \alpha_1} = \alpha_1 - \mathcal{K} \alpha_1 = 0$$

und

$$(45) \quad \mathcal{J}_k = \frac{\partial T_1}{\partial \alpha_k} = -\mathcal{K} \alpha_k = 0 \quad (k \neq 1).$$

Dies kann zusammengefaßt werden zu

$$(46) \quad \mathcal{J} = \frac{\partial T_1}{\partial \alpha} = \tilde{\alpha}_1 - \mathcal{K} \alpha = 0, \text{ wobei } ³⁶⁾,$$

und zwar (n, p) -reihig,

$$\tilde{\alpha}_1 = \begin{bmatrix} \alpha_1 & 0 & \dots & 0 \end{bmatrix}.$$

Aus (44) folgt, daß

$$(44') \quad \mathcal{K} \alpha_1 = \alpha_1.$$

Werden beide Seiten transponiert, so ist ³⁷⁾

$$(47) \quad (\mathcal{K} \alpha_1)' = \alpha_1' \mathcal{K} = \alpha_1'.$$

35) Die Differentiation von Vektoren und Matrizen folgt formal unmittelbar aus der entsprechenden Anordnung der differenzierten Koordinaten bzw. Elemente. Vgl.

DUSCHEK, A. (1950), a.a.O., 2, Seite 165-166 sowie RAO, C.R. (1952) - Advanced Statistical Methods in Biometric Research, New York: Wiley & Sons (London: Chapman & Hall), Seite 21 (Fußnote).

36) Bei (44) und (45) ist auf der äußersten rechten Seite der n -dimensionale Nullvektor (als Spalte), bei (46) die (n, p) -reihige Nullmatrix.

37) Da $\mathcal{K}$ eine symmetrische Matrix ist, gilt $\mathcal{K} = \mathcal{K}'$.

Diese Umformung können wir verwenden, wenn (46) mit α_1' linksmultipliziert wird. Aus

$$(48) \quad \alpha_1' \tilde{a}_1 - \alpha_1' a = 0$$

entsteht wegen (47)

$$(49) \quad \alpha_1' \tilde{a}_1 - \alpha_1' a = 0.$$

Nun sei (49) transponiert und umgestellt:

$$(50) \quad a' \alpha_1 - \tilde{a}_1' \alpha_1 = 0.$$

Durch Linksmultiplikation mit a und bei Beachtung von (36) erhalten wir

$$(51) \quad a \alpha_1 - a \tilde{a}_1' \alpha_1 = 0$$

mit

$$a \tilde{a}_1' \alpha_1 = \left(\sum_{i=1}^n a_{i1}^2 \right) \begin{bmatrix} a_{11} \\ a_{21} \\ - \\ a_{n1} \end{bmatrix}.$$

Definieren wir für (40) einen Skalar

$$(52) \quad \lambda = \sum_{i=1}^n a_{i1}^2 = \alpha_1' \alpha_1,$$

so geht aus (51)

$$(53) \quad a \alpha_1 - \lambda \alpha_1 = (a - \lambda \mathcal{E}) \alpha_1 = 0$$

hervor ³⁸⁾.

144 Das Ergebnis der Extremwertaufgabe

Die linke Seite von (53) ist eine $(n,1)$ -Matrix, eine Spalte, die ausgeschrieben die Form

38) Im Gegensatz zu (32) ist $\mathcal{E}$ hier die (n,n) -reihige Einheitsmatrix.

Grades in λ mit genau n Lösungen (Wurzeln) λ_h ($h=1,2,\dots$
 $\dots,n$), nämlich ⁴¹⁾

$$(54') \quad \text{Det}(\mathcal{R} - \lambda \mathcal{P}) = (-\lambda)^n + c_1(-\lambda)^{n-1} + \dots \\ \dots + c_{n-1}(-\lambda) + c_n = 0.$$

Blicken wir nun erneut auf das Maximumproblem (40), so hat uns unsere Ableitung auf das System (53') geführt, durch das der gesuchte Vektor α_1 nach Bestimmung von λ ermittelt werden kann. Da (40) und damit - wegen (52) - λ ein Maximum sein soll, ist offenbar unter den Lösungen von (54) der größte Wert herauszusuchen und für die Berechnung eines damit im Sinne der Komponentenanalyse optimalen Vektors α_1 zu verwenden.

Die Parameter λ und die zugehörigen Vektoren α_k stehen jedoch zur Korrelationsmatrix $\mathcal{R}$ in einer besonderen Beziehung, die vor Wiederaufnahme der Ableitung im nächsten Abschnitt näher zu untersuchen ist.

15 DIE HAUPTKOMPONENTENLÖSUNG ALS EIGENWERTPROBLEM DER KORRELATIONSMATRIX

150 Überführung in das Eigenwertproblem (Erste Hauptkomponente)

Die Gleichung (53) läßt sich auf die Form

$$(53'') \quad \mathcal{R} \alpha_1 = \lambda \alpha_1$$

41) Die Bedeutung der Koeffizienten c_i hat uns an dieser Stelle nicht weiter zu interessieren. Sie werden aus bestimmten Unterdeterminanten (Hauptminoren) der linken Seite von (54) gebildet, wobei speziell $c_1 = \sum_{i=1}^n r_{ii}$ und $c_n = \text{Det} \mathcal{R}$. Vgl. GANTMACHER, F.R. (1958), a.a.O., Seite 66-67. Bei der Zahl der Lösungen λ_h für (54') ist zu beachten, daß eventuell auftretende Mehrfachwurzeln entsprechend ihrer Vielfachheit zu zählen sind. Wir kommen auf dieses Problem auf Seite 38 zurück.

bringen. Ein solches Problem wird in der Theorie der Matrizen als Eigenwertaufgabe bezeichnet. Darunter versteht man die Aufgabe, zu einer quadratischen Matrix $\mathcal{R}$ einen Vektor u_1 derart zu suchen, daß der Vektor $\mathcal{R}u_1$ zu u_1 mit einem Faktor λ proportional ist. Ein solcher Vektor u_1 heißt Eigenvektor von $\mathcal{R}$, die aus (54) zu bestimmenden Zahlen λ_n werden Eigenwerte oder charakteristische Zahlen von $\mathcal{R}$ genannt. Die Matrix

$$(\mathcal{R} - \lambda \mathcal{I}) = \begin{bmatrix} r_{11} - \lambda & r_{12} & \dots & r_{1n} \\ r_{21} & r_{22} - \lambda & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \dots & r_{nn} - \lambda \end{bmatrix}$$

trägt den Namen charakteristische Matrix von $\mathcal{R}$, (54) heißt charakteristische Gleichung oder Säkulargleichung.

Die Lösung der Extremwertaufgabe hat mit (53) möglicherweise auf die Eigenwerte und Eigenvektoren von $\mathcal{R}$ geführt. Wenn wir nun den dazu erforderlichen Nachweis erbringen, so können wir das Problem der Komponentenanalyse als Eigenwertaufgabe der Korrelationsmatrix weiter verfolgen und uns dabei insbesondere der speziellen Eigenschaften der Eigenwerte und Eigenvektoren einer reellen, symmetrischen Matrix - wie sie, wie wir wissen, durch $\mathcal{R}$ gegeben ist - bedienen.

Die Vorbedingung dafür, daß wir es bei u_1 und λ in der Tat mit einem Eigenvektor und einem Eigenwert von $\mathcal{R}$ zu tun haben, ist in der charakteristischen Gleichung (54) zu suchen. Diese kann nämlich nur dann erfüllt werden, wenn die charakteristische Matrix $(\mathcal{R} - \lambda \mathcal{I})$ der Koeffizienten singulär ist⁴²⁾. Betrachten wir diese Frage in der Gleichung (53), so ist sicher, daß $\text{Det}(\mathcal{R} - \lambda \mathcal{I})$ nur dann ungleich Null sein könnte, wenn $u_1 = 0$ wäre. Diese triviale Lösung des Systems (53') kann aber deshalb nicht vorkommen, weil nach Voraussetzung alle

42) Diese Feststellung ist identisch mit der bereits auf Seite 30 erwähnten Bedingung, daß $\text{Det}(\mathcal{R} - \lambda \mathcal{I}) = 0$ sein muß.

Vektoren $u_k \neq 0$ sein müssen. Wäre das nämlich nicht der Fall, so könnte die Zahl der Komponenten im Widerspruch zu unseren Annahmen kleiner als p sein. Deshalb muß es in unserem Problem Zahlen λ so geben, daß $(R - \lambda \mathcal{P})$ unabhängig von singulärer oder nichtsingulärer Matrix R stets singulär mit $\text{Det } (R - \lambda \mathcal{P}) = 0$ ist.

Damit kann gesagt werden, daß $\lambda = S_1 = a_1' a_1$ ein Eigenwert und u_1 ein zugehöriger Eigenvektor der Matrix R ist. Wegen (40) ist zur Bestimmung der ersten Hauptkomponenten z_1 der größte unter den Eigenwerten von R auszusuchen. Zur Frage des Maximums ist weiter zu bemerken, daß auf eine Bildung der zweiten Ableitungen von (42) schon deshalb verzichtet werden kann, weil die Varianzen der $\mathcal{P}_i$ und der z_k Quadratsummen, also nicht-negativ sind ⁴³⁾.

151 Die Eigenwerte der Korrelationsmatrix

Die (n, n) -Matrix R ist reell und symmetrisch. Sie gehört damit zu der Klasse der diagonalähnlichen Matrizen, deren Eigenwerte und Eigenvektoren durch eine besonders einfache Struktur ausgezeichnet sind. Wir wollen jetzt untersuchen, wie dieser Sachverhalt im Sinne der Komponentenanalyse zu interpretieren ist.

Wir beginnen mit den n Eigenwerten und stellen fest, daß diese bei einer reell-symmetrischen Matrix stets reell und bei Zutreffen einer weiteren Bedingung sämtlich nicht-negativ sind. Beide Erscheinungen sind offenbar unerlässlich für die Komponentenanalyse, da wir als Norm des Vektors u_1 weder mit einer imaginären noch mit einer negativen Zahl sinnvoll arbeiten können. Für einen Beweis reeller Eigenwerte sei auf die Literatur verwiesen ⁴⁴⁾, auf die zweite Eigenschaft

43) Siehe dazu (38) und (39) auf Seite 24.

44) NEISS, F. (1955), a.a.O., Seite 101-102.

$$(55) \quad \lambda_h \geq 0 \quad (h=1, 2, \dots, n)$$

wollen wir im Hinblick auf weitere Anwendungen ausführlicher eingehen. Die Eigenwerte λ_h ⁴⁵⁾ folgen der Beziehung (55) dann, wenn $\mathcal{R}$ eine positiv definite Matrix ist. Wir erklären diesen Begriff dadurch, daß wir $\mathcal{R}$ als Formmatrix einer reellen quadratischen Form ⁴⁶⁾

$$(56) \quad Q = a'_k \mathcal{R} a_k$$

auffassen. Daraus folgt wegen (36)

$$(57) \quad Q = a'_k a a'_k = (a'_k a) (a'_k a)'$$

und damit

$$(56') \quad Q = (a'_k a_1)^2 + \dots + (a'_k a_k)^2 + \dots + (a'_k a_p)^2.$$

45) Hierzu noch eine Bemerkung. Die folgende Betrachtung könnte unnötig erscheinen deshalb, weil $\lambda = a'_1 a_1$ als eine Summe von Quadraten - vgl. (52) - nicht negativ sein kann. Es geht aber hier darum, für die Eigenwerte von $\mathcal{R}$ die Gültigkeit von (55) zu zeigen und damit nachzuweisen, daß der nicht-negative Skalar $a'_1 a_1$ ein Eigenwert von $\mathcal{R}$ sein kann.

46) Eine reelle quadratische Form in n reellen Veränderlichen a_{ik} ist eine in den a_{ik} homogene Funktion zweiten Grades mit den Koeffizienten r_{ih} , die durch

$$Q = a_{1k} (r_{11} a_{1k} + \dots + r_{1n} a_{nk}) + a_{2k} (r_{21} a_{1k} + \dots + r_{2n} a_{nk}) + \dots + a_{nk} (r_{n1} a_{1k} + \dots + r_{nn} a_{nk})$$

$$= \sum_{i,h=1}^n r_{ih} a_{ik} a_{hk} \text{ dargestellt werden kann, woraus sich}$$

die Schreibweise (56) ergibt. Die gemischten Glieder $a_{ik} a_{hk}$ ($i \neq h$) kommen je zweimal vor, man kann sie zu $(r_{ih} + r_{hi}) a_{ik} a_{hk}$ zusammenfassen und - da nur ihre Summe interessiert - $r_{ih} = r_{hi}$ setzen; das entspricht auch den Gegebenheiten unserer Analyse. Das Ergebnis der Form ist eine reelle Zahl. Wir werden später zu einer bedeutsamen geometrischen Anwendung quadratischer Formen zurückkehren. Siehe hierzu

GANTMACHER, F.R. (1958), a.a.O., Seite 270,

ZURMÜHL, R. (1958), a.a.O., Seite 127 und

AMATO, V. (1958) - Forme Quadratiche in Statistica. Mailand: A. Giuffrè, Seite 9-13.

Nun heißt eine reelle quadratische Form positiv definit, wenn sie nur eine positive Zahl oder den Wert Null ergeben kann. Da für uns $\alpha_k \neq 0$ gilt, muß bei allen Vektoren α_k in (56') wenigstens ein Summand - nämlich das Quadrat des inneren Produktes $\alpha_k' \alpha_k$ - stets positiv sein ⁴⁷⁾. Deshalb gilt

$$(56'') \quad Q > 0 \text{ für jeden Vektor } \alpha_k \quad (k=1,2,\dots,p).$$

Eine reelle quadratische Form wird durch ihre Koeffizienten, in unserem Falle durch die Elemente der Matrix $\mathcal{R}$, eindeutig festgelegt. Deshalb identifiziert man Form und Matrix insoweit, als mit positiv definiter Form Q auch die zugehörige Matrix $\mathcal{R}$ positiv definit genannt wird.

Wir wollen jetzt von p Eigenwerten für $\mathcal{R}$ ausgehen und (53'') auf die Form

$$(58) \quad \mathcal{R} \alpha_k = \lambda_k \alpha_k \quad (k=1,2,\dots,p)$$

bringen. Wird (58) mit α_k' von links multipliziert, so gilt

$$(59) \quad \alpha_k' \mathcal{R} \alpha_k = Q = \lambda_k \alpha_k' \alpha_k.$$

Da $\alpha_k' \alpha_k \neq 0$ und positiv, kann man daraus den Ausdruck ⁴⁸⁾

$$(60) \quad \lambda_k = \frac{\alpha_k' \mathcal{R} \alpha_k}{\alpha_k' \alpha_k} > 0 \quad (k=1,2,\dots,p)$$

bilden. Damit konnten wir aus dem Ansatz (56) heraus zeigen, daß die Eigenwerte λ_k positiv sein müssen. Das ist aber unbefriedigend, weil $\mathcal{R}$ genau n Eigenwerte λ_k besitzt und daher (60) nur für $(p=n)$, also $\lambda_k = \lambda_n$, von Interesse ist.

Die Matrix $\mathcal{A}$ hat p Spalten. Sollen nur diese als Eigenvektoren von $\mathcal{R}$ - mit zugehörigen Eigenwerten λ_k - in Frage

47) Wir werden später sehen, daß in der Tat alle übrigen Summanden gleich Null sind.

48) (60) entspricht dem sogenannten RAYLEIGHSCHEM Quotienten $R(\alpha_k)$. Vgl. ZURMÜHL, R. (1958), a.a.O., Seite 182.

kommen, so muß gezeigt werden, daß im Falle ($p < n$) die Matrix $\mathcal{R}$ den $(n-p)$ -fachen Eigenwert $\lambda_{p+1} = \lambda_{p+2} = \dots = \lambda_n = 0$ besitzt. Hierzu müssen wir für Q und $\mathcal{R}$ eine weitere Unterscheidung vornehmen. Man nennt $\mathcal{R}$ eigentlich positiv definit, falls in Bezug auf

$$(61) \quad Q \geq 0$$

die Form Q nur für $\alpha_k = 0$ den Wert Null ergibt; dagegen ist $\mathcal{R}$ positiv semidefinit, wenn (61) auch bei $\alpha_k \neq 0$ zu dem Ergebnis Null führen kann ⁴⁹⁾. Dabei ist $\mathcal{R}$ im ersteren Falle nichtsingulär mit dem Rang ($p=n$) ⁵⁰⁾. Diese Möglichkeit brauchen wir nicht weiter zu untersuchen, da sie uns bereits durch (60) auf p reelle, positive Eigenwerte λ_k gewiesen hat. Die Vektoren α_k sind sämtlich ungleich Null und Q gehorcht der Beziehung (56''). Die Korrelationsmatrix ist andererseits semidefinit, wenn sie singulär mit dem Rangabfall ($d=n-p$) ist. Wir haben dann von einer Form ⁵¹⁾

$$(62) \quad Q = \alpha'_h \mathcal{R} \alpha_h \quad (h=1, 2, \dots, p, p+1, \dots, n)$$

auszugehen, die (61) statt (56'') erfüllt, also $Q=0$ auch bei $\alpha_h \neq 0$ ergeben kann. Wegen der nun vorliegenden linearen Abhängigkeit der α_h müssen dann aber gerade die $\lambda_{p+1}, \lambda_{p+2}, \dots, \lambda_n$ zu $Q=0$ führen; deswegen ist zu (60) der Quotient

$$(60') \quad \lambda_{p+m} = \frac{\alpha'_{p+m} \mathcal{R} \alpha_{p+m}}{\alpha'_{p+m} \alpha_{p+m}} = 0 \quad (p+m=p+1, p+2, \dots, n)$$

hinzuzufügen. Wir kehren nun zur Beziehung (55) zurück und stellen zusammenfassend fest, daß $\mathcal{R}$ p reelle, positive Eigenwerte

$$(63) \quad \lambda_k > 0 \quad (k=1, 2, \dots, p)$$

49) ZURMÜHL, R. (1958), a.a.O., Seite 131-132.

50) Vgl. unsere Ausführungen auf Seite 21.

51) Um die Einführung neuer Symbole zu vermeiden, erweitern wir vorübergehend die Matrix $\mathcal{A}$ um $(n-p)$ auf n Spalten.

besitzt, zu denen, wenn $\mathcal{R}$ positiv semidefinit ist, (n-p)-mal ein Eigenwert Null tritt ⁵²⁾. Da wir es vorläufig nur mit dem größten Eigenwert von $\mathcal{R}$ zu tun haben, können wir durch

$$(64) \quad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p > 0$$

eine Ordnung nach der Größe vereinbaren, die (53'') in die endgültige Form

$$(53''') \quad \mathcal{R} \alpha_1 = \lambda_1 \alpha_1$$

überführt.

152 Eigenvektoren, mehrfache Eigenwerte und System der Eigenachsen

Nachdem wir die speziellen Eigenschaften der Eigenwerte von $\mathcal{R}$ untersucht haben, müssen wir uns nun den zugehörigen Eigenvektoren zuwenden und nachprüfen, welche Bedeutung die Gleichsetzung derselben mit den Spalten α_k der Matrix $\mathcal{A}$ hat. Im Zusammenhang damit haben wir uns kurz mit der Frage des Auftretens positiver Mehrfachwurzeln (63) auseinanderzusetzen.

Zunächst sei daran erinnert, daß das unter 144 beschriebene Gleichungssystem (53') homogen in den Koordinaten des Vektors α_1 ist und daher nach Einsetzen von λ_1 eine Reihe von gleichberechtigten, aber nicht eindeutigen Lösungen liefern kann ⁵³⁾. Die Frage mehrdeutiger Lösungsvektoren ist in unserem Falle leicht zu klären. (53''') zeigt, daß lediglich das Verhältnis der Koordinaten von α_1 zueinander festzuliegen braucht; die Gleichung wird von jedem beliebigen Vielfachen dieses Verhältnisses erfüllt. Liegt ein linear unabhängiger Lösungsvektor vor - wir haben schon angedeutet, daß es auch mehrere sein können -, so existiert dazu eine

52) ZURMÜHL, R. (1958), a.a.O., Seite 183.

53) FEIGL-ROHRBACH (1953), a.a.O., Seite 163.

Schar von linear abhängigen Vektoren, die ein Vektorgebilde der ersten Dimension, eine Eigenrichtung, bilden. Diese Mehrdeutigkeit wird gewöhnlich dadurch beseitigt, daß der Eigenvektor auf die Länge 1 normiert wird. In unserem Falle haben wir es als Betrag oder Länge der α_k mit dem Ausdruck

$$(65) \quad |\alpha_k| = \sqrt{(\alpha_k, \alpha_k)} = \sqrt{\sum_{i=1}^n a_{ik}^2} = \sqrt{\lambda_k}$$

zu tun. Wir fassen also den Eigenvektor α_1 als Lösung der Eigenwertaufgabe mit λ_1 als eindeutig in dem Sinne auf, daß er auf die Länge $\sqrt{\lambda_1}$ normiert ist.

Wir kommen nun auf die Möglichkeit zurück, daß (53') zu einem Eigenwert λ_1 mehrere linear unabhängige Lösungsvektoren α_1 besitzt. Das tritt immer dann ein, wenn die charakteristische Matrix $(\mathcal{R} - \lambda \mathcal{P})$ einen Rangabfall von $(d > 1)$ hat⁵⁴⁾. Es existieren dann zu λ_1 ($d > 1$) linear unabhängige Lösungsvektoren, die einen $(d > 1)$ -dimensionalen Raum, den sogenannten Eigenraum, aufspannen. Das Auftreten solcher Eigenräume ist mit dem mehrfachen Eigenwerte verbunden⁵⁵⁾. Es gehören nämlich zu jedem einfachen Eigenwert ein linear unabhängiger Eigenvektor, zu einem mehrfachen Eigenwert der Vielfachheit $(s \leq p)$ ein Eigenraum, dessen Dimension wenigstens 1 und höchstens s beträgt. Eine spezielle Eigenschaft diagonalähnlicher Matrizen, zu denen $\mathcal{R}$ als reell-symmetrische Matrix gehört, ist es, daß zu ihnen genau n linear unabhängige Eigenvektoren existieren⁵⁶⁾. Daraus erhellt, daß die Dimension der Eigenräume als übereinstimmend mit der

54) Der Rangabfall ($d=1$) führt zu einem linear unabhängigen Lösungsvektor, also auf die bereits besprochene Eigenrichtung.

55) ZURMÜHL, R. (1958), a.a.O., Seite 154.

56) Hier sind, wie bei den Eigenwerten, die $(n-p)$ zu dem $(n-p)$ -fachen Eigenwert 0 gehörenden Vektoren eines $(n-p)$ -dimensionalen Eigenraumes mitgezählt. Es handelt sich dabei nicht um Nullvektoren! Sie bleiben in der Analyse unberücksichtigt, da α p -spaltig definiert ist.

Vielfachheit s genau festliegt. Wenn also dieser Fall unter den Eigenwerten von $\mathcal{R}$ vorkommt, so ist die Dimension der Eigenräume bekannt und es können innerhalb derselben willkürlich linear unabhängige Vektoren ausgewählt und als Spalten von $\mathcal{A}$ verwendet werden ⁵⁷⁾.

Wir haben noch zu erklären, daß bei reell-symmetrischer Matrix die Realität der Eigenwerte die Folge hat, daß auch die Eigenvektoren stets mit reellen Koordinaten dargestellt werden können. Zur Untersuchung weiterer Zusammenhänge ⁵⁸⁾ sei nun (58) für zwei verschiedene Eigenwerte $\lambda_k \neq \lambda_m$ betrachtet:

$$(66) \quad \mathcal{R}a_k = \lambda_k a_k; \quad \mathcal{R}a_m = \lambda_m a_m.$$

Wir multiplizieren beide Gleichungen wechselseitig mit a'_m und a'_k von links, nämlich

$$(67) \quad a'_m \mathcal{R}a_k = \lambda_k a'_m a_k; \quad a'_k \mathcal{R}a_m = \lambda_m a'_k a_m,$$

und transponieren die erste Gleichung:

$$(68) \quad a'_k \mathcal{R}a_m = \lambda_k a'_k a_m.$$

Durch Subtraktion erhalten wir

$$(69) \quad (\lambda_k - \lambda_m) a'_k a_m = 0.$$

Das führt wegen $\lambda_k \neq \lambda_m$ zu

57) Ist etwa $\lambda_1 = \lambda_2$ (d.h. $s=2$) und sind alle anderen Eigenwerte voneinander verschieden, so gibt es zu $\mathcal{R}$ ($p-2$) Eigenrichtungen und eine Eigenebene, in der die linear unabhängigen Eigenvektoren a_1 und a_2 frei auswählbar sind. Allerdings muß diese Auswahl - wie wir auf Seite 40 sehen werden - so geschehen, daß die Vektoren zueinander orthogonal sind. In der Literatur zur Komponentenanalyse wird übereinstimmend versichert, daß dieser Fall nur sehr selten eintritt. Das können wir nicht beurteilen, sehen aber selbst in seinem Auftreten einen Umstand, der mit der Wahl geeigneter Vektoren unschwer übergangen werden kann.

58) Zur Realität und Orthogonalität der Eigenvektoren siehe ZURMÜHL, R. (1958), a.a.O., Seite 184.

$$(70) \quad \alpha'_k \alpha_m = 0 \quad (k, m=1, 2, \dots, p; k \neq m).$$

Diese Beziehung kann mit (65) zu

$$(71) \quad \alpha'_k \alpha_m = \lambda_k \delta_{km} = \begin{cases} \lambda_k & \text{für } k=m \\ 0 & \text{für } k \neq m \end{cases}$$

zusammengefaßt werden⁵⁹⁾. Die Eigenvektoren von $\mathcal{R}$ bilden demnach ein System von p orthogonalen Vektoren mit fest gewählter Länge $\sqrt{\lambda_k}$.

Wir untersuchen nun die Gesamtheit der Eigenvektoren von $\mathcal{R}$. Es ist gezeigt worden, daß eine spaltenweise Zusammenstellung der Eigenvektoren bei ($p=n$) auf die bereits früher definierte Matrix $\mathcal{A}$ führt, die deshalb auch Eigenvektor- oder Modalmatrix von $\mathcal{R}$ genannt werden kann. Da nun die Eigenvektoren ein p -fach orthogonales Vektorgebilde darstellen, liegt es nahe, sie als p -dimensionale Orthogonalbasis aufzufassen. Dem entspricht ein rechtwinkliges Achsensystem im T_p - das System der Eigenachsen -, in dem möglicherweise die Beziehungen der Matrix $\mathcal{R}$ in besonders einfacher Form erscheinen. Die Eigenwerte λ_k können wegen (52) und (71) in einer (p, p)-reihigen Matrix untergebracht werden, die als Folge der Orthogonalität der Eigenvektoren eine Diagonalmatrix der Form

$$\alpha' \alpha = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix}$$

ist. Damit finden wir als Matrizengleichung für alle Beziehungen (58) den Ausdruck

$$(72) \quad \mathcal{R} \alpha = \alpha (\alpha' \alpha).$$

59) Hierzu sei auf die Gleichung (56') auf Seite 34 verwiesen. Da die Eigenvektoren orthogonal sind, ergibt Q ($p-1$)mal den Wert 0 und einmal die positive Zahl $(\sqrt{\lambda_k})^2 = \lambda_k$.

Bei ($p=n$) ist α quadratisch mit dem Rang p , die Matrix ist also nichtsingulär und besitzt eine inverse Matrix α^{-1} . Diese finden wir unter Zuhilfenahme der inversen Matrix von $\alpha'\alpha$. Die Diagonalmatrix hat sicher den Rang p und daher eine inverse Matrix

$$(\alpha'\alpha)^{-1} = \begin{bmatrix} \frac{1}{\lambda_1} & 0 & \dots & 0 \\ 0 & \frac{1}{\lambda_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\lambda_p} \end{bmatrix},$$

die die Bedingung

$$(73) \quad (\alpha'\alpha)^{-1}(\alpha'\alpha) = \mathcal{I}$$

erfüllt. Wird (73) mit der für ($p=n$) existierenden Matrix α^{-1} auf beiden Seiten von rechts multipliziert, so erhält man

$$(74) \quad (\alpha'\alpha)^{-1}\alpha' = \alpha^{-1}.$$

Diese Gleichung werden wir im Abschnitt 170 benötigen.

Wenn jetzt (72) mit α^{-1} linksmultipliziert wird, so ist

$$(75) \quad \alpha^{-1}\mathcal{R}\alpha = \alpha'\alpha.$$

Diesen Vorgang nennt man eine Ähnlichkeitstransformation⁶⁰⁾ auf das System der Eigenachsen. Die Matrizen $\mathcal{R}$ und $\alpha'\alpha$ heißen einander ähnlich, und zwar wegen der besonderen Gestalt von $\alpha'\alpha$ diagonalähnlich.

Leider können wir in unserer Analyse von der Beziehung (75) nur beschränkten Gebrauch machen; die Matrix $\mathcal{R}$ ist zwar stets diagonalähnlich, ähnlich zu $\alpha'\alpha$ nach der Definition von α jedoch nur im Falle ($p=n$), während wir auch auf die Möglichkeit ($p<n$) Rücksicht nehmen müssen. Dann nämlich ist α singulär und (72) kann nicht nach $\alpha'\alpha$ aufgelöst werden; mit anderen Worten ausgedrückt ist α nur eine unvollständige

60) ZURMÜHL, R. (1958), a.a.O., Seite 53-54 und 165-166.

Modalmatrix und für den T_p kommen nur p der n Eigenachsen - also ein Teilraum des von den Eigenvektoren aufgespannten Raumes - in Frage.

16 DURCHFÜHRUNG DER HAUPTKOMONENTENLÖSUNG

160 Portführung des Eigenwertproblems (Zweite Hauptkomponente)

In den Abschnitten 14 und 15 haben wir uns damit beschäftigt, die Koordinaten der Ausgangsvektoren $\mathcal{C}_i$ in Bezug auf einen Basisvektor $\mathcal{Z}_1$ so zu bestimmen, daß $\mathcal{Z}_1$ als Hauptkomponente einen maximalen Anteil an der Gesamtvariation der $\mathcal{C}_i$ erhält⁶¹⁾. Diese Koordinaten sind nach (12) und (28) die des ersten Spaltenvektors α_1 der Matrix $\mathcal{A}$. Die Analyse hat uns darauf geführt, daß der Varianzanteil (40) der ersten Hauptkomponente dem größten Eigenwert λ_1 der Korrelationsmatrix $\mathcal{R}$ entspricht und daß wir für α_1 die Koordinaten des zu λ_1 gehörenden Eigenvektors von $\mathcal{R}$ zu wählen haben. Es ist sicher, daß wir als nächstes die Koordinaten der $\mathcal{C}_i$ in Bezug auf die zweite Achse des neuen Koordinatensystems $\mathcal{Z}_1, \mathcal{Z}_2, \dots, \mathcal{Z}_p$ - die zweite Hauptkomponente - aufsuchen müssen, wobei als Bedingung zu stellen ist, daß $\mathcal{Z}_2$ zu $\mathcal{Z}_1$ orthogonal ist und daß der Anteil von $\mathcal{Z}_2$ an der Residualvariation der $\mathcal{C}_i$ maximiert wird. Es besteht aus der bisherigen Ableitung die Vermutung, daß die weitere Untersuchung auf den zweitgrößten Eigenwert λ_2 von $\mathcal{R}$ mit α_2 als zugehörigem Eigenvektor führen wird, jedoch bedarf dies eines gesonderten Beweises.

Wir gehen abermals von

$$(28'') \quad \mathcal{X} = \mathcal{A} \mathcal{Z}$$

aus und zerlegen $\mathcal{A}$ in zwei (n,p) -Matrizen, nämlich die uns bereits bekannte Matrix

$$\tilde{\mathcal{A}}_1 = [\alpha_1 \quad 0 \quad \dots \quad 0]$$

61) Vgl. Seite 20-21.

sowie als Ergänzung dazu

$$\tilde{\alpha}_{.1} = \begin{bmatrix} 0 & \alpha_2 & \dots & \alpha_p \end{bmatrix}.$$

Dann gilt

$$(76) \quad \mathcal{H} = (\tilde{\alpha}_1 + \tilde{\alpha}_{.1})z = \tilde{\alpha}_1 z + \tilde{\alpha}_{.1} z,$$

woraus

$$(77) \quad \mathcal{H} - \tilde{\alpha}_1 z = \tilde{\alpha}_{.1} z$$

folgt. Hierbei ist

$$\mathcal{H} - \tilde{\alpha}_1 z = \begin{bmatrix} x_{11} - a_{11}z_{11} & x_{12} - a_{11}z_{12} & \dots & x_{1N} - a_{11}z_{1N} \\ \hline x_{n1} - a_{n1}z_{11} & x_{n2} - a_{n1}z_{12} & \dots & x_{nN} - a_{n1}z_{1N} \end{bmatrix}$$

eine (n,N) -Matrix, die als Elemente die Ausgangswerte x_{ij} ($i=1,2,\dots,n; j=1,2,\dots,N$) vermindert um den Einfluß der ersten Hauptkomponenten z_1 , also abzüglich des ersten Summanden von (12') enthält. Wir wollen zur Vereinfachung eine Matrix

$$(78) \quad \mathcal{H}^*_{.1} = \mathcal{H} - \tilde{\alpha}_1 z$$

definieren, deren Elemente durch die Relation

$$(79) \quad x^*_{ij.1} = x_{ij} - a_{i1}z_{1j} \quad (i=1,2,\dots,n; j=1,2,\dots,N)$$

bestimmt werden. Dann haben wir in Analogie zu (12') und (28'') die Beziehungen

$$(80) \quad c^*_{i.1} = \sum_{k=2}^p a_{ik} z_k \quad (i=1,2,\dots,n)$$

sowie

$$(81) \quad \mathcal{H}^*_{.1} = \tilde{\alpha}_{.1} z.$$

In Übereinstimmung mit den Ausführungen unter 141 erhalten wir nun mit

$$\begin{aligned}
 (82) \quad \frac{1}{N} \mathcal{R}_{.1}^* \mathcal{R}_{.1}^{*'} &= \frac{1}{N} (\tilde{\alpha}_{.1} \mathcal{Z}) (\tilde{\alpha}_{.1} \mathcal{Z})' \\
 &= \frac{1}{N} \tilde{\alpha}_{.1} \mathcal{Z} \mathcal{Z}' \tilde{\alpha}_{.1}' \\
 &= \tilde{\alpha}_{.1} \tilde{\alpha}_{.1}'
 \end{aligned}$$

eine (n,n)-Matrix

$$(82') \quad \mathcal{R}_{.1}^* = \tilde{\alpha}_{.1} \tilde{\alpha}_{.1}' = \begin{bmatrix} \sum_{k=2}^p a_{1k}^2 & \sum_{k=2}^p a_{1k} a_{2k} & \dots & \sum_{k=2}^p a_{1k} a_{nk} \\ \text{---} & \text{---} & \text{---} & \text{---} \\ \sum_{k=2}^p a_{nk} a_{1k} & \sum_{k=2}^p a_{nk} a_{2k} & \dots & \sum_{k=2}^p a_{nk}^2 \end{bmatrix}$$

Zur Erklärung der Matrix $\mathcal{R}_{.1}^*$, die im Unterschied zu $\mathcal{R}^{(62)}$ keine Korrelationsmatrix für die Korrelationen zwischen den Vektoren $\mathcal{C}_{i.1}^*$ und $\mathcal{C}_{h.1}^*$ ist ⁶³⁾, sei festgestellt, daß die aus (79) entstandenen Variablen $x_{i.1}^*$ nicht in unserem Sinne normiert sind. Für sie gilt zwar

$$(83) \quad \bar{x}_{i.1}^* = 0,$$

dagegen ist

$$(83') \quad \text{Var } x_{i.1}^* = 1 + \frac{1}{N} a_{i1}^2 \quad (i=1,2,\dots,n).$$

Daher haben sie auch als Vektoren nicht wie die $\mathcal{C}_i$ nach (17) eine Länge von $\sqrt{N}$, sondern von

$$(84) \quad |\mathcal{C}_{i.1}^*| = \sqrt{(\mathcal{C}_{i.1}^*, \mathcal{C}_{i.1}^*)} = \sqrt{N(1 + a_{i1}^2)}.$$

$\mathcal{R}_{.1}^*$ ist also weder eine Korrelationsmatrix, noch eine Varianz-Kovarianz-Matrix ⁶⁴⁾ für die $\mathcal{C}_{i.1}^*$, sondern eine aus $\mathcal{R}$ nach Elimination der Auswirkung der ersten Hauptkomponenten entstandene Residualmatrix, für deren Elemente entsprechend zu (37)

62) Vgl. (30) auf Seite 18 und (36) auf Seite 23.

63) $\mathcal{R}_{.1}^*$ wird oft in mißverständlicher Weise als "matrix of residual correlations" bezeichnet.

64) Vgl. Fußnote 23 auf Seite 18.

$$(85) \quad r_{ih.1} = r_{ih} - a_{i1} a_{h1} = \begin{cases} \sum_{k=2}^p a_{ik}^2 & \text{für } i=h \\ \sum_{k=2}^p a_{ik} a_{hk} & \text{für } i \neq h \end{cases} \\ (i, h=1, 2, \dots, n)$$

gilt.

Für die Residualvarianz der $\mathcal{P}_i$ finden wir in der Hauptdiagonalen von $\mathcal{R}_{.1}^*$

$$(86) \quad \text{Res Var } \mathcal{P}_i = \text{Var } \mathcal{P}_i - a_{i1}^2 = \sum_{k=2}^p a_{ik}^2$$

und daraus für die zweite Hauptkomponente

$$(87) \quad S_2 = \sum_{i=1}^n a_{i2}^2 = \alpha_2' \alpha_2 = \text{Max} !$$

Wie in Abschnitt 142 sind bei der Maximierung von (87) die Nebenbedingungen

$$(88) \quad r_{ih.1} = \sum_{k=2}^p a_{ik} a_{hk} \quad (i, h=1, 2, \dots, n)$$

zu beachten.

Der Differentiationsprozeß vollzieht sich in völliger Analogie zu 143 und führt bei entsprechender Wahl der Symbole nach (52) und (53) mit

$$(87') \quad \lambda = \sum_{i=1}^n a_{i2}^2 = \alpha_2' \alpha_2$$

zu

$$(89) \quad \mathcal{R}_{.1}^* \alpha_2 - \lambda \alpha_2 = (\mathcal{R}_{.1}^* - \lambda \mathcal{E}) \alpha_2 = 0.$$

Hierbei ist für λ der größte unter den Eigenwerten von $\mathcal{R}_{.1}^*$ auszuwählen. Durch Einsetzen desselben in (89) kann der gesuchte Vektor α_2 - als zu λ gehöriger Eigenvektor von $\mathcal{R}_{.1}^*$ - berechnet werden.

161 Identifizierung von Eigenwerten und Eigenvektoren
nachgeordneter Residualmatrizen: Geometrie der
Hauptkomponentenlösung

Unsere Ausführungen zur zweiten Hauptkomponenten z_2 sind insoweit noch unvollständig, als wir mit (89) auf Eigenwerte und Eigenvektoren von $R_{.1}^*$ verwiesen wurden. Wir müssen sehen, ob dieses Ergebnis sich so umgestalten läßt, daß wieder die Eigenwerte und Eigenvektoren der Korrelationsmatrix R zum Zuge kommen, damit in (87') oder (89) λ durch λ_2 als zweitgrößtem Eigenwert von R ersetzt werden kann.

Hierfür bilden wir aus (85) ⁶⁵⁾

$$(90) \quad R_{.1}^* = R - a_1 a_1'$$

und erhalten in (89)

$$(91) \quad R a_2 - a_1 a_1' a_2 - \lambda a_2 = 0.$$

Aus (70) wissen wir, daß

$$(70') \quad a_1' a_2 = 0.$$

Daher gilt

$$(92) \quad R a_2 - \lambda a_2 = (R - \lambda E) a_2 = 0.$$

Wir dürfen also fortan (89) dahingehend interpretieren, daß die Extremwertaufgabe (87) auf einen Eigenwert von R geführt hat. Dabei bleibt zu untersuchen, ob tatsächlich der durch (89) gegebene größte Eigenwert von $R_{.1}^*$ auch der zweitgrößte Eigenwert λ_2 - im Sinne von (64) - von R ist. Hierzu zerlegen wir (36) in

$$(36') \quad R = a_1 a_1' + a_2 a_2' + \dots + a_p a_p'$$

und (82') mit (90) in

$$(90') \quad R_{.1}^* = R - a_1 a_1' = a_2 a_2' + a_3 a_3' + \dots + a_p a_p'.$$

65) Man beachte, daß $a_1' a_1$ eine reelle Zahl, dagegen $a_1 a_1'$ eine (n,n) -reihige Matrix ist.

Wenn nun, wie wir in Abschnitt 150 gesehen hatten, die Vektoren u_k die zu den positiven Eigenwerten von R gehörigen Eigenvektoren sind, so gilt nach (90'), daß alle Eigenwerte und Eigenvektoren von R bis auf λ_1 und u_1 auch Eigenwerte und Eigenvektoren von $R_{.1}^*$ sind. Daher ist der größte Eigenwert von $R_{.1}^*$, zu dem u_2 berechnet wird, der zweitgrößte Eigenwert von R .

Wir erwähnen schon an dieser Stelle, daß mit derselben Überlegung die Eigenwerte der bei der Koordinatenbestimmung für die $\mathfrak{z}_3, \mathfrak{z}_4, \dots, \mathfrak{z}_p$ analog zu (82') und (90) auftretenden Residualmatrizen $R_{.2}^*, R_{.3}^*, \dots, R_{.p-1}^*$ als Eigenwerte $\lambda_3, \lambda_4, \dots, \lambda_p$ - also als nach der Größe geordnete Eigenwerte - der Korrelationsmatrix R identifiziert werden können. Dies geschieht ohne Schwierigkeiten durch eine entsprechende Erweiterung der Beziehungen (36') und (90').

Der geometrische Zusammenhang zwischen den Vektoren $\mathcal{P}_1$ und den bei der Bestimmung ihrer Koordinaten in Bezug auf $\mathfrak{z}_2$ aufgetretenen Vektoren $\mathcal{P}_{i.1}^*$ bedarf einer näheren Erläuterung⁶⁶⁾. Wir haben nämlich in der Ableitung des Abschnitts 160 statt der Koordinaten der $\mathcal{P}_i$ eigentlich die der $\mathcal{P}_{i.1}^*$ in Bezug auf die zweite Hauptkomponente bestimmt. Das konnten wir unausgesprochen deshalb tun, weil beide Gruppen von Vektoren wegen (12') und (80) in Bezug auf die Basis $\mathfrak{z}_2, \mathfrak{z}_3, \dots, \mathfrak{z}_p$ dieselben Koordinaten haben. Nimmt man es genau, so haben wir den T_p in zwei weitere Teilräume $T_{(1)}$ und $T_{(p-1)}$ aufgespalten; dabei wird der $T_{(1)}$ von dem Vektor $\mathfrak{z}_1$ aufgespannt, er entspricht der "ersten Hauptrichtung" der $\mathcal{P}_{i.1}^*$ innerhalb des T_p . Im $T_{(p-1)}$ existieren als Basis der $\mathcal{P}_{i.1}^*$ die Vektoren $\mathfrak{z}_2, \mathfrak{z}_3, \dots, \mathfrak{z}_p$. Da die $\mathfrak{z}_k$ gegenseitig orthogonal sind, sind alle Vektoren im $T_{(1)}$ orthogonal zu denen

66) Die folgenden Ausführungen lehnen sich grundsätzlich an BOGAARD, P.J.M. VAN DEN; KAPTEIN, E. (1958) - Principal Components. Netherlands School of Economics: Econometric Institute, Report 5812, Seite 9-10, jedoch ist zu beachten, daß die Ausgangsvoraussetzungen dieser Autoren von den unseren verschieden sind.

im $T_{(p-1)}$ ⁶⁷⁾. Deshalb kann man sagen, daß die $\mathcal{P}_{i.1}^*$ die orthogonalen Projektionen ⁶⁸⁾ der $\mathcal{P}_i$ auf den $T_{(1)}$ sind. Die Maximierung des Anteils (87) von $\mathcal{J}_2$ an der Gesamtvariation hat daran anschließend zu den Koordinaten der $\mathcal{P}_{i.1}^*$ - und damit der $\mathcal{P}_i$ - in Bezug auf die Basis des $T_{(p-1)}$ geführt. So muß die Analyse mit der orthogonalen Projektion der $\mathcal{P}_{i.1}^*$ auf den $T_{(2)}$ zu den $\mathcal{P}_{i.2}^*$ führen und mit der Bestimmung ihrer Koordinaten in Bezug auf den $T_{(p-2)}$, und so fort, ihren weiteren Gang nehmen.

162 Allgemeine Lösung (k-te Hauptkomponente)

Wir wollen nun dazu übergehen, die Festlegung der Koordinaten der $\mathcal{P}_i$ in Bezug auf einen Basisvektor $\mathcal{J}_k$ in allgemeiner Form darzustellen. Gesucht ist also die Spalte α_k der Matrix $\mathcal{A}$, wobei vorauszusetzen ist, daß die Spalten $\alpha_1, \alpha_2, \dots, \alpha_{k-1}$ bereits bekannt sind. Da wir uns in den vorhergehenden Abschnitten bereits ausführlich mit der Bestimmung der Koordinaten α_1 und α_2 für die erste und zweite Hauptkomponente auseinandergesetzt haben, können wir uns in der Verallgemeinerung kurz fassen.

Aus (28'') erhalten wir ⁶⁹⁾ bei Zerlegung von $\mathcal{A}$ in

$$\tilde{\alpha}_{k-1} = \begin{bmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_{k-1} & 0 & \dots & 0 \end{bmatrix}$$

sowie

$$\tilde{\alpha}_{.k-1} = \begin{bmatrix} 0 & 0 & \dots & \alpha_k & \alpha_{k+1} & \dots & \alpha_p \end{bmatrix}$$

-
- 67) In einem solchen Fall heißt der Teilraum $T_{(1)}$ orthogonales Komplement von $T_{(p-1)}$. Vgl. GANTMACHER, F.R. (1958), a.a.O., Seite 246.
- 68) Man kann sich leicht davon überzeugen, daß die Projektion eines Vektors wieder ein Vektor ist.
- 69) Die folgenden Gleichungen (93) bis (101) geben in geraffter Form die Gleichungen (76) bis (92), Seite 43-46, wieder, wobei stets davon ausgegangen wird, daß nicht $(k=2, 3, \dots, p)$, sondern $(m=k, k+1, \dots, p)$ Hauptkomponenten hinsichtlich ihrer Koordinaten noch nicht ermittelt wurden.

die Beziehungen

$$(93) \quad \mathcal{H} = \tilde{\alpha}_{k-1} \mathcal{Z} + \tilde{\alpha}_{k-1} \mathcal{Z}$$

und

$$(94) \quad \mathcal{H} - \tilde{\alpha}_{k-1} \mathcal{Z} = \tilde{\alpha}_{k-1} \mathcal{Z}.$$

Mit den Elementen

$$(95) \quad x_{ij.k-1}^* = x_{ij} - \sum_{m=1}^{k-1} a_{im} z_{mj} \quad (i=1,2,\dots,n; j=1,2,\dots,N)$$

bilden wir als Residuum der Ausgangsmatrix nach der Bestimmung von $(k-1)$ Hauptkomponenten die Matrix

$$(95') \quad \mathcal{H}_{.k-1}^* = \mathcal{H} - \tilde{\alpha}_{k-1} \mathcal{Z}$$

und bekommen als Vektorgleichung

$$(96) \quad \mathcal{C}_{i.k-1}^* = \sum_{m=k}^p a_{im} z_m \quad (i=1,2,\dots,n),$$

als Matrizengleichung aus (94) und (95')

$$(97) \quad \mathcal{H}_{.k-1}^* = \tilde{\alpha}_{k-1} \mathcal{Z}.$$

Die Residualmatrix $\mathcal{R}_{.k-1}^*$ mit

$$(98) \quad r_{ih.k-1} = r_{ih} - \sum_{m=1}^{k-1} a_{im} a_{hm} = \begin{cases} \sum_{m=k}^p a_{im}^2 & \text{für } i=h \\ \sum_{m=k}^p a_{im} a_{hm} & \text{für } i \neq h \end{cases}$$

(i=1,2,...,n)

erhält man durch

$$(99) \quad \frac{1}{N} \mathcal{H}_{.k-1}^* \mathcal{H}_{.k-1}^{*'} = \frac{1}{N} \tilde{\alpha}_{k-1} \mathcal{Z} \mathcal{Z}' \tilde{\alpha}_{k-1}'$$

$$= \tilde{\alpha}_{k-1} \tilde{\alpha}_{k-1}' = \mathcal{R}_{.k-1}^*.$$

Zu maximieren ist

$$(99'') \quad S_k = \sum_{i=1}^n a_{ik}^2 = \mathbf{a}_k' \mathbf{a}_k = \text{Max} !$$

unter den Nebenbedingungen

$$(98') \quad r_{ih.k-1} = \sum_{m=k}^p a_{im} a_{hm} \quad (i, h=1, 2, \dots, n).$$

Das führt mit

$$(100) \quad (\mathcal{R}_{.k-1}^* - \lambda \mathcal{E}) \alpha_k = 0,$$

wobei

$$(101) \quad \lambda = \sum_{i=1}^n a_{ik}^2 = \alpha_k' \alpha_k,$$

auf den größten Eigenwert der Matrix $\mathcal{R}_{.k-1}^*$, den wir nach entsprechender Anwendung der Gleichungen (90) bis (92) sowie (90') mit dem Eigenwert λ_k unter den nach der Größe geordneten Eigenwerten von $\mathcal{R}$ gleichsetzen können.

Wird dieser Vorgang p-mal durchgeführt, so sind in Gestalt der p Spalten von $\mathcal{A}$ alle Koordinaten der $\mathcal{E}_i$ in Bezug auf das Orthogonalsystem $\mathcal{Z}_k$ bekannt. Wir können dann die Vektoren $\mathcal{E}_i$ nach (12') aus ihrem ursprünglichen Koordinatensystem $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_N$ in das System $\mathcal{Z}_1, \mathcal{Z}_2, \dots, \mathcal{Z}_p$ überführen⁷⁰⁾ oder, anders ausgedrückt, die von uns eingangs gesuchten Variablen $\mathcal{Z}_k$ als lineare Funktionen der $\mathcal{X}_i$ aus dem Gleichungssystem (11) berechnen. Wie gezeigt worden ist, erhalten wir die Spalten α_k nach Bestimmung der Eigenwerte λ_k von $\mathcal{R}$ als Eigenvektoren der Korrelationsmatrix. $\mathcal{R}$ als reell-symmetrische Matrix mit dem Rang p besitzt genau p von Null verschiedene und positive Eigenwerte; nur die zu diesen gehörenden Eigenvektoren sind für $\mathcal{A}$ zu verwenden. Das entspricht unserer Aufgabenstellung insoweit, als wir die Dimension des ursprünglich n-dimensionalen Komplexes nach Möglichkeit verringern sollen. Selbst wenn das nicht durchführbar ist, bietet die Ausgestaltung der Hauptkomponenten die Gewähr, daß die Varianzen der $\mathcal{E}_i$ von weniger als n oder gar weniger als p der $\mathcal{Z}_k$ in ausreichender Weise ausgeschöpft werden. Zu dieser Frage wollen wir noch eine gesonderte Ableitung folgen lassen.

70) Vgl. dazu die Ausführungen unter 122 auf Seite 14.

163 Gesamtvariation und Eigenwerte der Korrelationsmatrix

Die Ausgangsvariablen x_i waren zu Beginn der Analyse so normiert worden, daß

$$(5') \quad \text{Var } x_i' = 1 \quad (i=1,2,\dots,n).$$

Mit Hilfe von (16) kann man daher für die Summe aller Varianzen den Ausdruck

$$(102) \quad \frac{1}{N} \sum_{i=1}^n x_i x_i' = n$$

setzen. Aus dem System (38) können wir folgern, daß die Summe der Varianzen weiterhin durch

$$(103) \quad \sum_{i=1}^n \alpha_i^* \alpha_i^{*'} = n$$

gegeben ist. Nun ergibt sich aus der Struktur von α , daß

$$(104) \quad \sum_{i=1}^n \alpha_i^* \alpha_i^{*'} = n = \sum_{k=1}^p \alpha_k' \alpha_k,$$

wobei wir in der rechten Seite die Summe der Anteile der Hauptkomponenten z_k an allen Varianzen der x_i , nämlich

$$(39''') \quad \sum_{k=1}^p s_k = \sum_{k=1}^p \alpha_k' \alpha_k$$

erkennen. Dafür können wir aber durch Verallgemeinerung von (52) und wegen (104)

$$(105) \quad \sum_{k=1}^p \lambda_k = n$$

schreiben. Die Theorie der Eigenwerte lehrt, daß die Summe der Eigenwerte einer Matrix gleich der Summe der Elemente in ihrer Hauptdiagonalen ist ⁷¹⁾. Demzufolge haben wir für die Korrelationsmatrix

71) ZURMÜHL, R. (1958), a.a.O., Seite 30 und 160. Die Summe der Elemente der Hauptdiagonalen heißt Spur der Matrix. Sie wird gewöhnlich mit "Sp" gekennzeichnet.

$$(106) \quad \text{Sp } \mathcal{R} = \sum_{i=1}^n r_{ii} = \sum_{i=1}^n \lambda_i = \sum_{k=1}^p \lambda_k + (n-p)0.$$

Aus (31) und (37'') wissen wir andererseits, daß

$$(107) \quad \sum_{i=1}^n r_{ii} = \sum_{i=1}^n \alpha_i^* \alpha_i^{*'} = n.$$

Damit sind wir wieder auf der linken Seite von (104) und können sagen, daß mit der Festlegung der Hauptkomponenten durch die positiven Eigenwerte von $\mathcal{R}$ die gesamte Variation der $\mathcal{C}_1$ erfaßt wird.

Der Anteilswert einer Hauptkomponenten $\mathfrak{z}_k$ an der Summe der Varianzen der $\mathcal{C}_1$ läßt sich nach (105) durch den Quotienten λ_k/n darstellen. Die Bildung dieses Quotienten, den man für praktische Zwecke am besten als Prozentzahl

$$\left(\frac{\lambda_k}{n} 100\right) \%$$

berechnet, nach Bestimmung einer jeden Hauptkomponenten bietet bei fortlaufender Aufsummierung die Möglichkeit, den wachsenden Grad der Erfassung der Varianzen zu verfolgen und bei Erreichen der von der Sachlage her zulässigen Untergrenze die Analyse abzubrechen. Wir werden später anhand konkreter Untersuchungen sehen, daß im allgemeinen die ersten zwei oder drei Hauptkomponenten mehr als 75 % der Gesamtvariation auf sich vereinigen und daß die folgenden $\mathfrak{z}_k$ in ihrem Anteil sehr rasch abnehmen, wie es im Zuge fortschreitender Maximierung der Residuen nicht anders zu erwarten ist.

17 DIE ERGEBNISSE DER HAUPTKOMONENTENLÖSUNG UND IHRE BEZIEHUNG ZUR HAUPTACHSENTTRANSFORMATION

170 Gewinnung der Hauptkomponenten

Die Matrix $\mathcal{A}$ ist jetzt vollständig bekannt. Wir können daher zur Bestimmung von $\mathfrak{z}$ aus (28), unserer eigentlichen Aufgabe, übergehen. Linksmultiplikation von (28) mit $\mathcal{A}'$

ergibt

$$(108) \quad \alpha'x = \alpha'az.$$

Die Matrix $\alpha'a$, die Diagonalmatrix der positiven Eigenwerte von $\mathcal{R}$, ist uns aus Abschnitt 152 bereits bekannt ⁷²⁾.

Wir wissen, daß sie eine inverse Matrix besitzt und können deshalb (108) wie folgt nach z auflösen:

$$(109) \quad z = (\alpha'a)^{-1}\alpha'x.$$

Ein Vergleich von (109) mit (74) zeigt, daß die auf den Fall ($p < n$) mit singulärer Matrix A und unvollständigem System der Eigenachsen zugeschnittene Herleitung der Hauptkomponenten durch (108) bei ($p=n$) direkt aus (28) folgt. Wir haben dann

$$(110) \quad z = a^{-1}x,$$

wobei (109) bei nichtsingulärem A wegen (74) in (110) übergeht.

Für die in der (p, N) -Matrix $\mathcal{Z}$ als Zeilen angeordneten Hauptkomponenten liefern (109) oder (110) das System

$$(111) \quad z_k = \frac{1}{\lambda_k} \alpha'_k x \quad (k=1, 2, \dots, p),$$

das man als Umkehrung unserer ursprünglichen Formulierung (11) in der Form

$$\begin{aligned} z_1 &= \frac{1}{\lambda_1} (a_{11}x_1 + a_{21}x_2 + \dots + a_{n1}x_n) \\ (111') \quad z_2 &= \frac{1}{\lambda_2} (a_{12}x_1 + a_{22}x_2 + \dots + a_{n2}x_n) \\ &\quad \text{-----} \\ z_p &= \frac{1}{\lambda_p} (a_{1p}x_1 + a_{2p}x_2 + \dots + a_{np}x_n) \end{aligned}$$

oder

$$(111'') \quad z_k = \frac{1}{\lambda_k} \sum_{i=1}^n a_{ik}x_i \quad (k=1, 2, \dots, p)$$

schreiben kann.

72) Vgl. Seite 40.

Den Beziehungen (11') und (111') ist zu entnehmen, daß nicht nur die $\mathcal{C}_i$ als Linearkombinationen der $\mathcal{Z}_k$, sondern auch die $\mathcal{Z}_k$ als Linearkombinationen der $\mathcal{C}_i$ dargestellt werden können. Wir wollen noch anmerken, daß die Elemente der Vektoren $\mathcal{Z}_k$, die aus (111) berechnet werden können, ihre Koordinaten in Bezug auf die zu Beginn der Analyse beschriebene N-dimensionale Orthonormalbasis des V_N sind ⁷³⁾.

171 Nachweis der Eigenschaften der Hauptkomponenten

Im Abschnitt 120 war die Forderung, daß die Hauptkomponenten normiert und unkorreliert sein sollen ⁷⁴⁾, durch

$$(20') \quad (\mathcal{Z}_k, \mathcal{Z}_m) = N \delta_{km} = \begin{cases} N & \text{für } k=m \\ 0 & \text{für } k \neq m \end{cases}$$

ausgedrückt worden. Die Vektoren $\mathcal{Z}_k$ sollen demzufolge paarweise orthogonal sein und eine Länge von $\sqrt{N}$ haben. Wir wollen prüfen, ob sie dies in der Form (111) erfüllen.

Der Beweis sei in der Ausdrucksweise des Matrizenkalküls vorgenommen. Wir greifen dazu auf

$$(32') \quad \frac{1}{N} \mathcal{Z} \mathcal{Z}' = \mathcal{E}$$

als "Korrelationsmatrix" der Vektoren $\mathcal{Z}_k$ zurück. Wir haben dann wegen (109)

$$(112) \quad \begin{aligned} \frac{1}{N} \mathcal{Z} \mathcal{Z}' &= \frac{1}{N} [(\alpha' \alpha)^{-1} \alpha' \mathcal{X}] [(\alpha' \alpha)^{-1} \alpha' \mathcal{X}]' \\ &= \frac{1}{N} (\alpha' \alpha)^{-1} \alpha' \mathcal{X} \mathcal{X}' \alpha [(\alpha' \alpha)^{-1}]', \end{aligned}$$

wobei aus (30) und (36)

$$(113) \quad \frac{1}{N} \mathcal{X} \mathcal{X}' = \mathcal{R} = \alpha \alpha'$$

folgt und durch Symmetrie der Diagonalmatrix

73) Vgl. Abschnitt 122, Seite 14-15.

74) Vgl. (6) und (10), Seite 4 und 5.

$$(114) \quad [(a'a)^{-1}]' = (a'a)^{-1}$$

ist. Deshalb führt (112) zu

$$(115) \quad \frac{1}{N} \{ \{ \}' = (a'a)^{-1} (a'a) (a'a) (a'a)^{-1}.$$

Daraus haben wir nach (73) ⁷⁵⁾

$$(116) \quad \frac{1}{N} \{ \{ \}' = \mathcal{E}.$$

Auch diese Ableitung kann bei ($p=n$) durch die Anwendung von (110) vereinfacht werden.

172 Komponentenanalyse und Hauptachsentransformation

Bei der Untersuchung der verschiedenen Ausprägungen der Komponentenanalyse in der Literatur, die wir uns für einen späteren Abschnitt vorgenommen haben, wird man häufig auf ein Theorem der linearen Algebra geführt, das als Reduktion einer symmetrischen Matrix auf Diagonalgestalt bezeichnet wird und bei seiner geometrischen Anwendung auf quadratische Formen den Namen Hauptachsentransformation trägt. Daher wollen wir uns jetzt der Aufgabe unterziehen, dieses Theorem darzustellen und seine Beziehungen zur Komponentenanalyse, wie wir sie bisher formuliert haben, zu überprüfen.

Wir gehen wieder von der Eigenwertaufgabe nach (58) und (72) mit der Gleichung

$$(117) \quad R \bar{a} = \bar{a} \Lambda$$

aus. Dabei soll R die uns bekannte Korrelationsmatrix sein, $\bar{a}$ aber im Gegensatz zu a die Modalmatrix mit allen n linear

75) In (115) befindet sich die rechte Seite von (73) multipliziert mit sich selbst in umgekehrter Reihenfolge der Faktoren, so daß (116) das Produkt von zwei p -reihigen Einheitsmatrizen enthält. $\mathcal{E}$ kann jedoch bei gleicher Reihenzahl durch $\mathcal{E}$ ersetzt werden.

unabhängigen Eigenvektoren ⁷⁶⁾ $\bar{\alpha}_g$ ($g=1,2,\dots,n$); diese als Spalten zusammengestellt und durch die Beziehung

$$(118) \quad \bar{\alpha}'_f \bar{\alpha}_g = \delta_{fg} \quad (f,g=1,2,\dots,n)$$

normiert ⁷⁷⁾. $\bar{\alpha}$ ist also eine nichtsinguläre Matrix mit n Spaltenvektoren, die wegen (118) Einheitsvektoren sind. Eine solche Matrix heißt orthogonal ⁷⁸⁾, man kann zeigen, daß für sie

$$(119) \quad \bar{\alpha}^{-1} = \bar{\alpha}'$$

gilt. Schließlich wollen wir in (117) unter Λ die vollständige Eigenwertmatrix

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}$$

verstehen ⁷⁹⁾.

Nun spricht man von einer Hauptachsentransformation, wenn (117) unter Beachtung von (119) nach Λ aufgelöst und damit $\mathcal{R}$ durch Ähnlichkeitstransformation ⁸⁰⁾

$$(120) \quad \bar{\alpha}' \mathcal{R} \bar{\alpha} = \Lambda$$

auf Diagonalgestalt gebracht wird ⁸¹⁾. Dieses Ergebnis ist

76) Es sei daran erinnert, daß $\mathcal{R}$ gemäß unseren Feststellungen auf Seite 38 genau n linear unabhängige Eigenvektoren besitzt, unbeeinflusst vom Rangabfall d von $\mathcal{R}$. Vgl. auch Fußnote 56 auf Seite 38.

77) (118) steht im Widerspruch zu (71), Seite 40. Jedoch sind die Eigenvektoren, soweit sie linear unabhängig sind (Eigenrichtungen), nach ihrer Länge frei wählbar.

78) ZURMÜHL, R. (1958), a.a.O., Seite 30-31.

79) Hierbei ist es entgegen $\bar{\alpha}'\bar{\alpha}$ auf Seite 41 nicht ausgeschlossen, daß verschiedene der $\lambda_g = 0$ sind. Wie wir wissen, kommt das dann vor, wenn $\mathcal{R}$ singulär ist.

80) Vgl. (75) auf Seite 41.

81) SPERNER, E. (1957), a.a.O., 2, Seite 146; THRALL-TORNHEIM (1957), a.a.O., Seite 155-156 und Seite 188; GANTMACHER, F. R. (1958), a.a.O., Seite 283-284; AMATO, V. (1958), a.a.O., Seite 37-43; ZURMÜHL, R. (1958), a.a.O., Seite 190.

uns nicht unbekannt, wir haben es in Abschnitt 152 bereits diskutiert. Der Unterschied zu der dort vorgenommenen Betrachtung besteht allerdings darin, daß wir es hier durch Verwendung der orthogonalen Matrix $\bar{A}$ in jedem Falle mit dem stets n-dimensionalen System der Eigenachsen zu tun haben, das überdies wegen (118) orthonormal ist.

Übertragen wir die Hauptachsentransformation auf eine der Beziehung (117) entsprechende reelle quadratische Form, so ist diese gegenüber (60) und (62) mit

$$(121) \quad F = \bar{a}'_g R \bar{a}_g = \sum_{i,h=1}^n r_{ih} \bar{a}_{ig} \bar{a}_{hg}$$

anzusetzen. Angesichts der Annahme ($p=n$) ist F eigentlich positiv definit, aus (56) und (118) läßt sich darüber hinaus folgern, daß

$$(121') \quad \bar{a}'_g R \bar{a}_g = 1 \quad (g=1,2,\dots,n).$$

In geometrischer Anschauung stellt (121') die Mittelpunkts-gleichung einer Hyperfläche 2. Grades (Quadrik) im n-dimensionalen Raum dar. Eine beliebige Transformation

$$(122) \quad \bar{a}_g = \bar{A} w_g$$

überführt diese nach (120) in die Hauptachsenform

$$\begin{aligned} (123) \quad F &= w'_g \bar{A}' R \bar{A} w_g \\ &= w'_g \Lambda w_g \\ &= \lambda_1 v_{1g}^2 + \lambda_2 v_{2g}^2 + \dots + \lambda_n v_{ng}^2 = \sum_{h=1}^n \lambda_h v_{hg}^2 = 1. \end{aligned}$$

Aus (123) läßt sich ersehen, daß es sich um ein Hyperellipsoid ⁸²⁾ handelt, dessen Hauptachsen mit den Eigenachsen der

82) Siehe hierzu

CRAMÉR, H. (1946) - Mathematical Methods of Statistics. Princeton: Princeton University Press, Seite 114-115.
GROTEMEYER, K.P. (1958) - Analytische Geometrie (Sammlung Götschen, Bd. 65/65a). Berlin: Walter de Gruyter & Co., Seite 100-103 und Seite 113-122.
ZURMÜHL, R. (1958), a.a.O., Seite 190-191.

Matrix $\mathcal{R}$ zusammenfallen und dessen Halbachsenquadrate zu den Eigenwerten λ_n umgekehrt proportional sind ⁸³⁾.

Die hier geschilderte Beziehung zwischen der Eigenwertangabe und den Hauptachsen einer Hyperfläche 2. Grades erfährt im Falle ($p < n$) eine Abänderung dahingehend, daß in (123) wegen

$$(124) \quad \lambda_{p+1} = \lambda_{p+2} = \dots = \lambda_n = 0$$

die Glieder $v_{pg}^2, v_{p+1,g}^2, \dots, v_{ng}^2$ fortfallen. Damit ist F positiv semidefinit und das Ellipsoid verwandelt sich in einen elliptischen Zylinder ⁸⁴⁾, dessen Gestalt von der Größe der Differenz ($n-p$) abhängt.

Wir wollen noch anmerken, daß bei Annahme eines festliegenden, aber nicht notwendigerweise normalen n -dimensionalen Verteilungsgesetzes für die $x_1, x_2, \dots, x_n$ mit konstanter Wahrscheinlichkeitsdichte das Hyperellipsoid (123) dem von CRAMÉR beschriebenen Konzentrationsellipsoid ⁸⁵⁾ (ellipsoid of concentration) entspricht. Voraussetzung ist dazu allerdings, daß die mehrdimensionale Wahrscheinlichkeitsverteilung der dann als Zufallsvariable aufzufassenden x_i nichtsingulär ($p=n$) ist.

173 Spezielle Deutung der Hauptachsentransformation

Die geometrische Analogie der Hauptachsentransformation läßt sich auf die von uns dargestellte Form der Komponentenanalyse nicht übertragen, und zwar weder bei nichtsingulärer

83) Dies folgt aus einem Vergleich von (123) mit der Hauptachsengleichung
$$\frac{v_1^2}{A_1^2} + \frac{v_2^2}{A_2^2} + \dots + \frac{v_n^2}{A_n^2} = 1,$$

in der die A_h ($h=1,2,\dots,n$) den Halbachsen des Hyperellipsoids entsprechen.

84) GROTEMEYER, K.P. (1958), a.a.O., Seite 104, Gleichung 9.

85) CRAMÉR, H. (1946), a.a.O., Seite 300-301.

noch bei singulärer Korrelationsmatrix. Dies liegt in der Tatsache begründet, daß wir von Anfang an die Eigenvektoren u_k so definiert haben, daß die Matrix A wohl aus orthogonalen, nicht aber auf die Länge 1 normierten Eigenvektoren besteht⁸⁶⁾. Daher ist unsere Matrix A im Gegensatz zur vollständigen Modalmatrix $\bar{A}$ nicht orthogonal, ja im Falle ($p < n$) nicht einmal quadratisch. Zur geometrischen Veranschaulichung haben wir bei ($p = n$) anstelle von (121) von der quadratischen Form

$$(125) \quad Q = a'_k R a_k > 0 \quad (k=1, 2, \dots, p)$$

auszugehen. Diese ergibt aber wegen (59) und (60) nicht 1, sondern die positive Zahl

$$(126) \quad Q = \lambda_k^2$$

und damit erst durch

$$(127) \quad Q_1 = \frac{a'_k R a_k}{\lambda_k^2} = 1.$$

Hierbei führt die entsprechende Anwendung von (122), nämlich

$$(122') \quad a_k = a_{10k},$$

unter Verwendung von (72) zu

$$\begin{aligned} (128) \quad Q_1 &= \frac{a'_{10k} A' R A a_{10k}}{\lambda_k^2} \\ &= a'_{10k} \frac{1}{\lambda_k^2} (A' A) (A A) a_{10k} \\ &= a'_{10k} \varphi_{10k} \\ &= \sum_{h=1}^n v_{hk}^2 = 1 \quad (p=n). \end{aligned}$$

Dies ist aber im Gegensatz zu (123) die Mittelpunktsgleichung

86) Vgl. Seite 38 und Seite 40-41.

einer reellen Hyperkugel ⁸⁷⁾, deren Hauptachsen die gleiche Länge haben und eine Deutung der Eigenvektoren oder Eigenwerte nicht zulassen. HOTELLING, der in seiner Analyse den Weg des Abschnitts 172 einschlägt, hat gleichwohl unseren Fall in Erwägung gezogen und schreibt ⁸⁸⁾:

"If, instead of the y's or z's, the γ 's or other independent an equally variable quantities be taken as rectangular coordinates, the ellipsoids are squeezed and stretched into spheres. Each test is represented by a line through the origin. The correlation between two tests is the cosine of the angle between their lines" ⁸⁹⁾.

Für ($n < p$), also eine positiv semidefinite Form (125), ist A eine (n, p)-reihige, das heißt rechteckige Matrix. Die Beziehung (72) gilt auch hier, jedoch muß zur Transformation

$$(122'') \quad \alpha_k = A w_k^*$$

der Vektor w_k^* p -zeilig definiert werden. Damit haben wir

$$(129) \quad Q_2 = \sum_{m=1}^p v_{mk}^2 = 1 \quad (p < n).$$

Das heißt aber, daß hier die bei (123) zu beobachtende zylindrische Entartung der Hyperfläche nicht stattfindet. Vielmehr bleibt die Hyperkugel stets in der Dimension des Ranges von R erhalten.

87) SPERNER, E. (1957), a.a.O., 2, Seite 373-374.

88) HOTELLING, H. (1933), a.a.O., Seite 428.

89) Die " γ 's" entsprechen den z_k . Unter einem "test" ist eine Variable x_i zu verstehen (Anmerkung des Verf.).

2	<u>NUMERISCHE KOMPONENTENANALYSE</u>
20	DIE RECHENTECHNIK DER KOMPONENTENANALYSE
200	Allgemeines zur Rechentechnik

Der Sinn der bisherigen Ausführungen bestand darin, für die Methode der Komponentenanalyse ein einwandfreies theoretisches Fundament zu legen. Fragen der praktischen Durchführung einer konkreten Analyse wurden dabei - von einigen Ausnahmen abgesehen - nicht berührt. Diese Auffassung soll auch jetzt nicht aufgegeben werden, weil die numerische Behandlung der Komponentenanalyse weitgehend in das Gebiet der praktischen Rechentechnik für kompliziertere mathematische Beziehungen fällt, für das eine umfangreiche Fachliteratur existiert und das als solches nicht Gegenstand dieser Arbeit sein kann. Da wir aber vorhaben, die hier erarbeitete Methode an einem einfachen Zahlenbeispiel zu erläutern, erscheint es notwendig, diesem eine Reihe von vorbereitenden Ausführungen vorzuschicken.

Ein wichtiges Ergebnis der theoretischen Überlegungen des vorigen Kapitels bestand in dem Nachweis, daß das Problem der Hauptkomponenten unter geeigneten Voraussetzungen auf eine Form gebracht werden kann, in der die Eigenwerte und Eigenvektoren einer reellen, symmetrischen Matrix von entscheidender Bedeutung sind. In der Tat können die Hauptkomponenten β_k nach (111) nur dann bestimmt werden, wenn neben der ohnehin gegebenen Ausgangsmatrix $\mathcal{X}$ die Eigenwerte λ_k und die entsprechend normierten Eigenvektoren u_k der Korrelationsmatrix $\mathcal{R}$ bekannt sind. Wenn wir die Berechnung von $\mathcal{R}$ als formal einfaches und allenfalls bei größeren Werten für n und N durch den Rechenaufwand schwieriges Problem hier übergehen wollen, so besteht demnach die wesentliche Anforderung an eine konkrete Komponentenanalyse in der Auffindung der Eigenwerte und Eigenvektoren der Korrelationsmatrix. Dabei darf die verhältnismäßig einfache Struktur, die das Eigenwertproblem in der allgemeinen Formulierung der

Abschnitte 144 und 150 besitzt, nicht darüber hinwegtäuschen, daß dies bei Korrelationsmatrizen höherer Ordnung eine rechentechnische Aufgabe beträchtlichen Umfanges ist⁹⁰⁾. Fast immer bedarf diese der überlegten Anlage und sorgfältigen Durchführung, sollen Zeitaufwand und Schwierigkeitsgrad nicht in ein unangemessenes Verhältnis gegenüber dem Untersuchungszweck geraten.

Wir wollen noch erwähnen, daß die hier angedeuteten Schwierigkeiten die praktische Anwendung der Komponentenanalyse erschwert haben. Tatsächlich hat man in konkreten Fällen oft zugunsten weniger exakter, aber rechentechnisch leichter Methoden auf ihre Verwendung verzichtet. Dabei kommt häufig die dem Verfahrenskreis der Faktorenanalyse entstammende Schwerpunkts- oder Zentroidmethode (centroid method of factoring) zum Zuge, da diese - wie wir im Abschnitt 310 andeuten werden - unter geeigneten Voraussetzungen "Faktoren" liefert, die als gute Annäherungen an die Hauptkomponenten gelten können. So entnehmen wir einer Arbeit von BARGMANN die Bemerkung⁹¹⁾:

"Die mathematisch exakte Bestimmung dieses Vektors⁹²⁾ ist praktisch schwer durchführbar, denn sie erfordert die Auflösung sehr großer Säkulargleichungssysteme. ...Bisher sind nahezu alle Methoden zur Auffindung der 'exakten' Hauptachsenlösung an der ungeheuer langwierigen rechnerischen Arbeit gescheitert. Die beste Annäherung, die sogenannte 'Schwerpunktslösung', ist dagegen so einfach zu bestimmen, daß sie fast ausschließlich angewendet wurde."

Ähnliche Ansichten finden sich an anderen Stellen der Literatur; wir begnügen uns mit einem Hinweis auf BURT, der die

90) So bedenke man zum Beispiel, daß die charakteristische Gleichung (54) - vgl. Seite 30-31 - bereits bei 10 Ausgangsvariablen ($n=10$) eine algebraische Gleichung 10. Grades ist. Allerdings gibt es Verfahren, mit deren Hilfe die Auflösung einer solchen Gleichung umgangen werden kann.

91) BARGMANN, R. (1956) - "Überführung der Schwerpunkts- in die Hauptachsenlösung und weitere Entwicklung der Näherungsverfahren in der Faktorenanalyse", Mittbl. Math. Stat. 8, Seite 2.

92) Gemeint ist die erste Hauptkomponente (Anm.d.Verf.).

Komponentenanalyse als "far too laborious to employ in actual practice" bezeichnet ⁹³⁾.

Es unterliegt keinem Zweifel, daß der Wert eines wissenschaftlichen Verfahrens ungeachtet der Präzision seiner Aussagen von dem Grad der praktischen Anwendungsmöglichkeit mitbestimmt wird. Insoweit weist die hier betrachtete Methode der Komponentenanalyse einen gewissen Nachteil auf. Wir möchten jedoch behaupten, daß dieser Nachteil im Zeichen der Fortschritte der neuzeitlichen Rechenverfahren auf maschinell oder gar elektronischem Wege immer mehr ausgeglichen wird ⁹⁴⁾, so daß in absehbarer Zeit die Vorteile der Komponentenanalyse gegenüber anderen Verfahren, die vielfach als statistisch unbefriedigend angesehen werden müssen, weitaus stärker als bisher zur Geltung kommen.

201 Abschätzung der Eigenwerte; direkte Verfahren zur Eigenwertbestimmung

Bei den praktischen Methoden zur Lösung der Eigenwertaufgabe der Korrelationsmatrix R ist die Tatsache zu beachten,

93) BURT, C. - Diskussionsbeitrag zu:
BARTLETT, M.S. (1947) - "Multivariate analysis", J. Roy. Stat. Soc. Suppl. 2, Seite 192.

94) Siehe hierzu
HALL, D.M.; WELKER, E.L.; CRAWFORD, J. (1945) - "Factor analysis calculations by tabulating machines", Psychometrika 10, Seite 93-125.
HOUSEHOLDER, A.S. (1956) - "Numerical mathematics from the viewpoint of electronic digital computers", Nachrichtentechn. Fachber. 4, Seite 21-25.
Im übrigen darf man schon wegen der zahlreichen technischen Anwendungen vermuten, daß die Bestimmung der Eigenwerte und Eigenvektoren einer Matrix - also die numerische Hauptaufgabe der Komponentenanalyse - als Standardprogramm elektronisch gesteuerter Rechenanlagen geführt wird.

daß $\mathcal{R}$ stets reell und symmetrisch ist ⁹⁵⁾. Wir wissen aus unseren ausführlichen Betrachtungen des Abschnitts 151, daß aus diesem Grunde genau n reelle und nicht negative Eigenwerte existieren müssen. Der Bereich der möglichen Eigenwerte ist dadurch erheblich eingeschränkt und einer ersten Kontrolle unterwerfbar. Es gibt weiterhin eine Reihe von Möglichkeiten, den Wertebereich der

$$(55') \quad \lambda_h \geq 0 \quad (h=1,2,\dots,n)$$

im voraus zusätzlich einzugrenzen. Von diesen erwähnen wir hinsichtlich der Größe von λ_h die Beziehung ⁹⁶⁾

$$(130) \quad \lambda_h \leq n$$

sowie als weitere Einengung

$$(131) \quad 0 \leq \lambda_h \leq \max_i \sum_{h=1}^n |r_{ih}| = \max_h \sum_{i=1}^n |r_{ih}|,$$

wobei die rechte Seite die größte (nicht notwendig positive) Zeilen- und damit zugleich Spaltensumme von $\mathcal{R}$ wiedergibt ⁹⁷⁾.

- 95) Einen guten Überblick über die praktischen Methoden bieten die entsprechenden Kapitel bei
 ZURMÜHL, R. (1957²) - Praktische Mathematik für Ingenieure und Physiker. Berlin (Göttingen, Heidelberg): Springer-Verlag, das Kapitel VI in
 ZURMÜHL, R. (1958), a.a.O., Seite 273-343, sowie
 DWYER, P.S. (1951) - Linear Computations. New York: Wiley & Sons (London: Chapman & Hall) Seite 219-235.
 SCHOLZ, H. (1959) - Praktische Lösung von Aufgaben der Matrizenrechnung. In: Anwendungen der Matrizenrechnung auf wirtschaftliche und statistische Probleme (Einzelschr. d. Deutsch. Stat. Ges. Nr. 9), Würzburg: Physica-Verlag, Seite 235-262.
 Eine kurze Zusammenfassung findet man bei
 TINTNER, G. (1952) - Econometrics. New York: Wiley & Sons (London: Chapman & Hall), Seite 351-358.
- 96) Diese Abschätzungsformel nimmt nur im Falle einer Korrelationsmatrix eine so einfache Form an. Sie lautet eigentlich $|\lambda_h| = n(\max_{i,h} r_{ih})$, wobei der Wert λ_h wegen (55) nur positiv sein kann und das betragsmäßig größte Element stets 1 ist. Vgl. ZURMÜHL, R. (1958), a.a.O., Seite 203.
- 97) (131) ist eine Modifizierung der Gleichungen (30) und (31) bei ZURMÜHL, R. (1958), a.a.O., Seite 205.

Als natürlicher Weg zur Bestimmung der Eigenwerte bietet sich die Aufstellung und Lösung der charakteristischen Gleichung (54') an. Im Gegensatz dazu stehen Verfahren, die die Säkulargleichung vermeiden und die Eigenwerte durch iteratives Vorgehen gewinnen. Man unterscheidet demzufolge die direkte und die indirekte Eigenwertbestimmung. Beide Wege sind mit Vor- und Nachteilen verbunden, die je nach Art der Untersuchung und dem Vorhandensein technischer Hilfsmittel gegeneinander abzuwägen sind.

Der direkte Fall präsentiert sich als Lösung einer algebraischen Gleichung n-ten Grades, dem aber aus (54) die Aufstellung der Säkulargleichung vorauszugehen hat. Der übliche Weg der Auflösung der Determinante der charakteristischen Matrix ⁹⁸⁾ $(R - \lambda E)$ nach ihren Zeilen oder Spalten erfordert bei größerem Wert n einen beträchtlichen Rechenaufwand, der allenfalls durch gewisse Vereinfachungstechniken ⁹⁹⁾ verringert werden kann. Die Aufstellung der charakteristischen Gleichung empfiehlt sich deshalb nur bei Korrelationsmatrizen verhältnismäßig geringer Reihenzahl, wozu insbesondere das Verfahren von HESSENBERG ¹⁰⁰⁾ zur Verfügung steht. Ist die Gleichung

$$(54'') \quad \text{Det}(R - \lambda E) = 0$$

bekannt, so kann die Berechnung der n Eigenwerte nach den gewöhnlichen Methoden erfolgen ¹⁰¹⁾; Komplikationen durch

98) Vgl. Seite 31.

99) REIERSØL, O. (1940) - "A method for recurrent computation of all the principal minors of a determinant and its application in confluence analysis", Ann. Math. Stat. 11, Seite 193-198.
DWYER, P. S. (1941a) - "The evaluation of determinants", Psychometrika 6, Seite 191-204.
- - - Ders. (1951), a.a.O., Seite 141-165.

100) Auf dieses gehen wir hier nicht ein. Ausführliche Beschreibung bei
ZURMÜHL, R. (1957), a.a.O., Seite 138-146 und
- - - Ders. (1958), a.a.O., Seite 306-312.

101) Dabei sind das HORNER-Schema und das Verfahren von GRAEFTE von besonderem Interesse. Vgl.
ZURMÜHL, R. (1957), a.a.O., Seite 34-91.

imaginäre Wurzeln sind nicht zu erwarten. Daran anschließend sind die Eigenwerte λ_h nacheinander in das System (53') zur Bestimmung der Koordinaten der Eigenvektoren einzusetzen ¹⁰²⁾. Wir hatten bereits festgestellt ¹⁰³⁾, daß (53') ein lineares Gleichungssystem ist, das wegen seiner Homogenität n linear unabhängige und sogar orthogonale, aber nicht eindeutige Lösungsvektoren liefert. Diese Schwierigkeit kann dadurch umgangen werden, daß man willkürlich etwa die erste Koordinate eines jeden Eigenvektors gleich 1 setzt, die damit überbestimmten n Gleichungssysteme nach den verbleibenden (n-1) Koordinaten auflöst und anschließend eine Normierung nach (65) durchführt ¹⁰⁴⁾. Bezeichnen wir die so berechneten Eigenvektoren ¹⁰⁵⁾ mit

$$\alpha_k^o = \begin{bmatrix} a_{1k}^o \\ a_{2k}^o \\ \vdots \\ a_{nk}^o \end{bmatrix},$$

dann erfolgt die Normierung auf die Länge $\sqrt{\lambda_k}$ auf dem Wege

$$(132) \quad \alpha_k = \frac{\sqrt{\lambda_k}}{|\alpha_k^o|} \alpha_k^o \quad (k=1,2,\dots,p).$$

Aus diesen Vektoren α_k wird schließlich die Matrix α

102) Im Falle des HESSENBERG-Verfahrens können die Eigenvektoren auf einem etwas anderen Wege gewonnen werden.

103) Vgl. Seite 37-38.

104) Dieses Verfahren versagt, wenn die erste Koordinate des Eigenvektors den Wert 0 besitzt. In diesem Falle müßte man die Normierung nach einer anderen, nicht verschwindenden Koordinate unter den verbleibenden (n-1) Werten ausrichten.

105) Da die Eigenvektoren verschwindender Wurzeln von (54') nicht beachtet werden, können wir abermals das Symbol λ_h ($h=1,2,\dots,n$) durch λ_k ($k=1,2,\dots,p$) ersetzen.

zusammengesetzt, mit deren Hilfe man die gesuchten Vektoren $\mathbf{z}_k$ berechnen kann ¹⁰⁶⁾.

202 Indirekte Eigenwertbestimmung: Iterationsverfahren

Bei größeren Korrelationsmatrizen, wie sie in der Komponentenanalyse der Praxis die Regel sein werden, sind die indirekten oder iterativen Verfahren zur Eigenwertbestimmung vorteilhafter als die im vorigen Abschnitt beschriebene Aufstellung und Lösung der Säkulargleichung, zumal bei ihnen zumeist die Eigenvektoren simultan mit den Eigenwerten bestimmt werden können. Aus diesem Grunde haben die indirekten Verfahren in der Literatur zur Komponentenanalyse stets eine erhebliche Rolle gespielt, insbesondere ist auch HOTELLING mit entsprechenden Veröffentlichungen hervorgetreten.

Wir beginnen mit einer kurzen Darstellung des "klassischen" Iterationsverfahrens, das auf VON MISES zurückzuführen ist ¹⁰⁷⁾. Es besteht darin, einen beliebigen, reellen Spaltenvektor $\mathbf{b}_0$ zu wählen, diesen mit der Matrix $\mathbf{R}$ von links zu multiplizieren, den so gewonnenen Vektor $\mathbf{R}\mathbf{b}_0$ erneut mit $\mathbf{R}$ zu multiplizieren, und so fort. In allgemeiner Darstellung entspricht dieses Vorgehen der Vorschrift

$$(133) \quad \mathbf{b}_y = \mathbf{R}_{y-1} \quad (y=1, 2, \dots),$$

106) Aus einer ganzen Reihe von Arbeiten zur Praxis der Auflösung linearer Gleichungssysteme seien erwähnt: CHAUNCEY, H. (1939) - "A method for solving simultaneous equations, with particular reference to multiple correlation problems", Harv. Educ. Rev. 9, Seite 63-68. DWYER, P. S. (1941b) - "The solution of simultaneous equations", Psychometrika 6, Seite 101-129. FOX, L. (1950) - "Practical methods for the solution of linear equations and the inversion of matrices", J. Roy. Stat. Soc. B. 12, Seite 120-136. BURT, C. (1951) - "The numerical solution of linear equations", Brit. J. Psychol. Stat. Sect. 4, Seite 31-54.

107) ZURMÜHL, R. (1957), a.a.O., Seite 137-138, Seite 159-172. - - -Ders. (1958), a.a.O., Seite 179-180, Seite 273-303.

Nun kann man zeigen, daß sich mit zunehmender Iterationsstufe der Vektor b_y dem zum größten Eigenwert ¹⁰⁸⁾ λ_1 gehörenden Eigenvektor a_1^0 nähert, während das Verhältnis der Koordinaten zweier aufeinanderfolgender Vektoren, nämlich

$$(134) \quad q_{iy} = \frac{b_{iy}}{b_{i,y-1}} \quad (i=1,2,\dots,n),$$

gegen eben diesen Wert λ_1 konvergiert.

Die Begründung dieser Eigenschaft liegt darin, daß R als diagonalähnliche Matrix genau n linear unabhängige Eigenvektoren besitzt und b_0 daher als Linearkombination

$$(135) \quad b_0 = d_1 a_1^0 + d_2 a_2^0 + \dots + d_n a_n^0$$

darstellbar ist. Daraus folgt aber bei Iteration

$$(136) \quad b_1 = R b_0 = d_1 R a_1^0 + d_2 R a_2^0 + \dots + d_n R a_n^0,$$

wofür man wegen

$$(137) \quad R a_h^0 = \lambda_h a_h^0 \quad (h=1,2,\dots,n)$$

$$(138) \quad b_1 = d_1 \lambda_1 a_1^0 + d_2 \lambda_2 a_2^0 + \dots + d_n \lambda_n a_n^0$$

schreiben kann. Die Fortsetzung der Iteration führt zu

$$(138') \quad b_y = d_1 \lambda_1^y a_1^0 + d_2 \lambda_2^y a_2^0 + \dots + d_n \lambda_n^y a_n^0$$

oder

$$(138'') \quad b_y = \lambda_1^y \left[d_1 a_1^0 + d_2 \left(\frac{\lambda_2}{\lambda_1}\right)^y a_2^0 + \dots + d_n \left(\frac{\lambda_n}{\lambda_1}\right)^y a_n^0 \right].$$

Wenn wir nun annehmen, daß λ_1 betragsmäßig erheblich größer als die folgenden Eigenwerte und ($d_1 \neq 0$) ¹⁰⁹⁾ ist, dann

108) Vgl. (64) auf Seite 37.

109) Das bedeutet, daß der Ausgangsvektor b_0 wenigstens eine Koordinate des ersten Eigenvektors besitzen muß. Wir haben bereits vorgeschlagen, den Eigenvektor a_1^0 an erster -an sich beliebiger- Stelle mit der Zahl 1 zu versehen; die Bedingung ($d_1 \neq 0$) ist dann erfüllt, wenn man, wie es meistens geschieht, den Ausgangsvektor in der Form $b_0 = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}$ ansetzt.

verschwinden mit fortschreitender Iteration die Glieder der eckigen Klammer bis auf $d_1 \alpha_1^0$; es gilt schließlich ¹¹⁰⁾

$$(139) \quad \lim_{y \rightarrow \infty} \frac{\beta_y}{\lambda_1^y} = d_1 \alpha_1^0$$

und damit

$$(134') \quad \lim_{y \rightarrow \infty} \frac{b_{iy}}{b_{i,y-1}} = \lambda_1 \quad (i=1,2,\dots,n).$$

Aus diesen Ausführungen geht hervor, daß die Konvergenz des Verfahrens dann besonders gut ist, wenn λ_1 ein "dominanter", das heißt betragsmäßig stark überwiegender Eigenwert ist. Liegt das nicht vor, so können sich sehr langwierige Iterationsprozesse ergeben. Ein Nachteil der indirekten Methode besteht außerdem darin, daß die Bestimmung der auf λ_1 folgenden Eigenwerte besondere Zusatzmaßnahmen erfordert (Verfahren von KOCH), weil dann - etwa bei λ_2 - in (138'') ($d_1=0$) gelten muß, durch die Iteration aber immer wieder Einflüsse von α_1^0 bemerkbar werden und in regelmäßigen Abständen durch Orthogonalisierung von α_1^0 und β_y ausgeschaltet werden müssen. Auf Sonderfragen, die bei betragsgleichen oder verschwindenden Eigenwerten auftreten, wollen wir hier nicht weiter eingehen. Gegenüber den direkten Verfahren hat das iterative Vorgehen für die Rechenpraxis den bedeutenden Vorteil, daß Fehler, die sich in einen Vektor β_y einschleichen, im Zuge der Iteration von selbst bereinigt werden und allenfalls den Prozeß verlängern können. Außerdem ist durch geeignete Summenproben eine Kontrolle auf jeder Stufe möglich, schließlich dienen die Gleichungen (71) und (105) als wirksame Endkontrolle.

203 Ergänzungen zur Rechentechnik: Iterationsverfahren nach HOTELLING, Methode von TUCKER

Die indirekte Methode der Iteration wurde hier in ihrem Grundprinzip dargestellt. Durch vielfältige Verbesserungen läßt sie sich - wie AITKEN, RICHARDSON und AKAIKE gezeigt

¹¹⁰⁾ (139) und (134') sind Abänderungen der Gleichungen (8) und (9) bei ZURMÜHL, R. (1957), a.a.O., Seite 161.

haben ¹¹¹⁾ - auf eine Form bringen, die auch für die Zwecke der Komponentenanalyse geeigneter ist. HOTELLING beginnt seine Ausführungen zur iterativen Lösung wie folgt ¹¹²⁾:

"The explicit calculation of the determinant ω and its principal minors, and the solution of the characteristic equation and of the homogenous linear equations (17) ¹¹³⁾ would be a laborious computation. A vast saving of arithmetical effort is effected by the following iterative method, which yields simultaneously a root and the corresponding coefficients, the roots appearing in order of magnitude, the greatest first. This makes it possible to stop whenever it is evident that all the important principal components have been obtained".

Daran schließt sich die Diskussion eines Iterationsprozesses an, die vor allem mit geometrischen Argumenten geführt wird und ein Vorgehen liefert, das weitgehend dem von uns geschilderten Verfahren entspricht. Wenn man auch vermuten darf, daß der Fall sehr kleiner Quotienten zwischen aufeinanderfolgenden Eigenwerten von HOTELLING nicht genügend berücksichtigt wurde, so ist auf der anderen Seite der Vorschlag interessant und außerordentlich arbeitssparend, bei jedem neuen Eigenwert λ_n von der reduzierten

-
- 111) AITKEN, A.C. (1937) - "Studies in practical mathematics. II. The evaluation of the latent roots and latent vectors of a matrix", Proc. Roy. Soc. Edinb. 57, Seite 269-303.
KINCAID, W.M. (1947) - "Numerical methods for finding characteristic roots and vectors of matrices", Quart. Appl. Math. 5, Seite 320-345.
RICHARDSON, L.F. (1950a) - "A purification method for computing the latent roots of numerical matrices and some integrals of differential equations", Roy. Soc. Lond. Phil. Trans. A. 242, Seite 439-491.
- - - Ders. (1950b) - "A method for computing principal axes", Brit. J. Psychol. Stat. Sect. 3, Seite 16-20.
AKAIKE, H. (1958) - "On a computation method for eigenvalue problems and its application to statistical analysis", Ann. Inst. Stat. Math. 10, Seite 1-20.
- 112) HOTELLING, H. (1933), a.a.O., Seite 429.
- 113) Gemeint sind die charakteristische Determinante der Korrelationsmatrix und das System (53'), Seite 30 (Anm. d. Verf.).

Korrelationsmatrix $\mathcal{R}_{.h-1}^*$ auszugehen ¹¹⁴⁾. Man kann dann wieder einen größten Eigenwert bestimmen, weil - wie wir eingehend besprochen haben - der zweitgrößte Eigenwert von $\mathcal{R}_{.h}^*$ dem größten Eigenwert von $\mathcal{R}_{.h-1}^*$ entspricht. Auf diese Weise läßt sich die Schwierigkeit der iterativen Bestimmung höherer Eigenwerte der Korrelationsmatrix umgehen; die Matrix $\mathcal{R}_{.h-1}^*$ ist aus einer Beziehung der Form (98) unschwer herzustellen.

In einer späteren Arbeit ¹¹⁵⁾ geht HOTELLING davon aus, daß sich (133) auch durch

$$(133') \quad \mathcal{b}_y = \mathcal{R}^y \mathcal{b}_0 \quad (y=1,2,\dots)$$

ausdrücken läßt. Daraus wird gefolgert und an einem Beispiel gezeigt, daß sich der Iterationsprozeß durch die Verwendung einer Potenz $(\mathcal{R}^2)^t$ ($t=1,2,\dots$) anstelle von $\mathcal{R}$ erheblich beschleunigen läßt ¹¹⁶⁾. ZURMÜHL ¹¹⁷⁾ bringt jedoch

114) Diese Kritik wird auch von FRISCH vorgebracht, der die Methode von HOTELLING im ökonometrischen Zusammenhang der Konfluenzanalyse untersucht. Vgl. FRISCH, R. (1934) - Statistical Confluence Analysis by Means of complete Regression Systems. Oslo: Universitets Økonomiske Institutt, Publikasjon nr. 5, Seite 35. Weitere kritische Bemerkungen finden sich bei VAN DEN BOGAARD-KAPTEIN (1958), a.a.O., Seite 20. Es ist hinzuzufügen, daß auch fast alle Standardwerke der Faktoranalyse Erörterungen zu den hier geschilderten Rechenverfahren enthalten.

115) HOTELLING, H. (1936a) - "Simplified calculation of principal components", Psychometrika 1, Seite 27-35. Vgl. HORST, P. (1937) - "A method of factor analysis by means of which all coordinates of the factor matrix are given simultaneously", Psychometrika 2, Seite 225-236. HOUSEHOLDER, A.S.; YOUNG, G. (1938) - "Matrix approximation and latent roots", Am. Math. Monthly 45, Seite 165-171, insbesondere Seite 168-170.

116) Dabei ist zu beachten, daß das Potenzieren von $\mathcal{R}$ auf - sagen wir - $\mathcal{R}^8$ nichts anderes ist als achtmalige Iteration nach (133) mit einem "Versuchsvektor" $\mathcal{b}_0 = \{1 \ 1 \ \dots \ 1\}$.

117) ZURMÜHL, R. (1958), a.a.O., Seite 276.

hierzu den Nachweis, daß diese Methode nur bei kleineren Matrizen mit hoher Iterationsstufe für die Konvergenz vorteilhaft ist.

Wir beschließen diesen Abschnitt mit der Erwähnung einer Zusammenfassung verschiedener Methoden von HOTELLING sowie weiterer Vorschläge zur Berechnung der Hauptkomponenten von FLOOD und TUCKER ¹¹⁸⁾.

Während FLOOD ein modifiziertes direktes Verfahren empfiehlt, welches von der Tatsache Gebrauch macht, daß $\mathcal{R}$ der eigenen charakteristischen Gleichung genügt ¹¹⁹⁾, zeigt TUCKER, daß nach der Ermittlung der ersten Hauptkomponenten z_1 die folgende Erweiterung der Korrelationsmatrix $\mathcal{R}$ eine Matrix

$$\mathcal{R}_{II} = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} & a_{11}\sqrt{-\lambda_1} \\ r_{21} & r_{22} & \dots & r_{2n} & a_{21}\sqrt{-\lambda_1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{11}\sqrt{-\lambda_1} & a_{21}\sqrt{-\lambda_1} & \dots & a_{n1}\sqrt{-\lambda_1} & -\lambda_1 \end{bmatrix}$$

ergibt, deren erste Hauptkomponente $z_{II.1}$ der zweiten Hauptkomponente z_2 der Matrix $\mathcal{R}$ entspricht. Auch auf diese Art braucht stets nur der größte Eigenwert bestimmt zu werden, wobei die $\mathcal{R}_{II}, \mathcal{R}_{III}, \dots$ symmetrisch, aber nicht reell bleiben. Das hat allerdings für die Eigenwerte dieser Matrizen keine Folgen. Erweitert man nämlich auch die Matrix $\mathcal{A}$

118) HOTELLING, H. (1943) - "Some new methods in matrix calculation", Ann.Math.Stat. 14, Seite 1-34, insbesondere Seite 26-33.

FLOOD, M.M. (1940) - "A computational procedure for the method of principal components", Psychometrika 5, Seite 169-172.

TUCKER, L.R. (1944a) - "The determination of successive principal components without computation of tables of residual correlation coefficients", Psychometrika 9, Seite 149-153.

119) Vgl. GANTMACHER, F.R. (1958), a.a.O., Seite 77 (Satz von CAYLEY-HAMILTON).

um eine $(n+1)$ te Zeile $\begin{bmatrix} \sqrt{-\lambda_1} & 0 & \dots & 0 \end{bmatrix}$, so liefert die Diagonalmatrix $a'_{II} a_{II}$ einen Eigenwert ¹²⁰⁾

$$(140) \quad \lambda_{II.1} = \sum_{i=1}^n a_{i1}^2 + (\sqrt{-\lambda_1})^2 = \lambda_1 - \lambda_1 = 0,$$

während alle anderen Eigenwerte reell bleiben und der größte von diesen dem zweitgrößten von $\mathcal{R}$ entspricht.

21 EIN ZAHLENBEISPIEL

210 Ausgangsgesamtheit und Korrelationsmatrix

Das Ziel der folgenden Abschnitte ist es, die vollständige Durchrechnung eines Zahlenbeispiels zur Komponentenanalyse zu demonstrieren. Wir beschränken uns dabei darauf, die Arbeitsweise des Verfahrens an einer kleinen Ausgangsgesamtheit aufzuzeigen, weil der Rechenaufwand für ein konkretes wirtschaftswissenschaftliches Forschungsobjekt ohne Verwendung von größeren maschinellen Einrichtungen für den hier beabsichtigten Zweck zu umfangreich ist.

Für unsere Aufgabe legen wir in starker Vereinfachung ($n=4$) sowie ($N=6$) fest und benutzen als Ausgangsvariable die folgenden, völlig willkürlich gewählten Zahlenreihen ¹²¹⁾.

Tabelle 1: Frei gewählte Zahlenwerte einer Ausgangsgesamtheit mit $n=4$; $N=6$.

N \ n	Ausgangsvariable			
	X ₁	X ₂	X ₃	X ₄
(1)	(2)	(3)	(4)	(5)
1	15	8	154,923	5,106
2	16	7	138,868	1,913
3	21	5	152,419	21,901
4	18	3	216,572	3,487
5	20	2	125,931	13,537
6	19	2	84,875	6,482

120) TUCKER, L.R. (1944a), a.a.O., Gleichungen (7) und (8) auf Seite 151.

121) Die Zahlen für X₃ und X₄ sind im Gegensatz zu X₁ und X₂ ohne besonderen Grund mit Dezimalstellen gewählt worden. Durch verschiedene Größenordnungen und Genauigkeitsangaben soll lediglich ein möglichst praxisnaher Eindruck vermittelt werden.

Im Zusammenhang mit den einleitenden Ausführungen des Abschnitts 110 wäre es möglich, die sechs Elemente etwa als Landkreise A, B, ..., F aufzufassen. Die Variablen X_i ($i=1, 2, 3, 4$) könnten dann als die für jeden Landkreis beobachteten Werte eines Merkmals aus der Wirtschaftsstatistik (zum Beispiel Beschäftigte, Steueraufkommen, gewerbliche Umsätze, Kraftfahrzeugbestand u.a.m.) betrachtet werden. Allerdings ist hinzuzufügen, daß derartige Berechnungen auf Grund so kleiner Ausgangsgesamtheiten nicht sinnvoll sein können; es ist deshalb nicht zweckmäßig, die angegebenen Variablen direkt mit konkreten Tatbeständen zu identifizieren.

Zur Durchführung der Transformationen

$$(4') \quad x_{ij} = \frac{1}{\sqrt{\text{Var } X_i}} (X_{ij} - \bar{X}_i) \quad (i=1, 2, 3, 4; j=1, 2, \dots, 6)$$

benötigen wir für jede Variable X_i den arithmetischen Mittelwert und die mittlere quadratische Abweichung.

Tabelle 2: Mittelwerte und mittlere quadratische Abweichungen der Ausgangsvariablen

n	$\bar{X}_i$	$\sqrt{\text{Var } X_i}$
(1)	(2)	(3)
X_1	18,16667	2,11476
X_2	5,58333	2,36291
X_3	145,59800	39,34920
X_4	19,88159	4,45888

Damit sind wir in der Lage, die Normierung vorzunehmen und nach der Vereinbarung des Abschnitts 130 die Matrix $\mathcal{H}$ aufzubauen.

$$\mathcal{H} = \begin{bmatrix} -1,49741 & -1,02455 & 1,33979 & -0,07881 & 0,86692 & 0,39405 \\ 1,48122 & 1,05802 & 0,21160 & -0,63481 & -1,05802 & -1,05802 \\ 0,23698 & -0,17101 & 0,17335 & 1,80370 & -0,49981 & -1,54318 \\ -0,47807 & -1,19417 & 1,27012 & -0,84117 & 1,41276 & 0,16948 \end{bmatrix}$$

Mit Hilfe von (8) lassen sich aus den in $\mathcal{H}$ als Zeilen angeordneten Variablen x_i ($i=1, 2, 3, 4$) die Korrelationskoeffizienten berechnen, die in der Korrelationsmatrix zusammengefaßt werden.

$$\mathcal{R} = \begin{bmatrix} 1 & r_{12} & r_{13} & r_{14} \\ r_{21} & 1 & r_{23} & r_{24} \\ r_{31} & r_{32} & 1 & r_{34} \\ r_{41} & r_{42} & r_{43} & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & -0,71709 & -0,18849 & 0,81089 \\ -0,71709 & 1 & 0,20388 & -0,41404 \\ -0,18849 & 0,20388 & 1 & -0,27512 \\ 0,81089 & -0,41404 & -0,27512 & 1 \end{bmatrix}$$

211 Die Berechnung der Eigenwerte

Die Abschätzungsformel (131) liefert für die gegebene Korrelationsmatrix den Wert

$$(131') \quad 0 \leq \lambda_h \leq \max_i \sum_{h=1}^4 |r_{ih}| = \max_h \sum_{i=1}^4 |r_{ih}| = 2,71647.$$

Die Bestimmung der Eigenwerte wird zunächst auf direktem Wege vorgenommen. Die charakteristische Gleichung (54') läßt sich bei (n=4) verhältnismäßig leicht aus der Determinante von $(\mathcal{R} - \lambda \mathcal{E})$ berechnen, weil diese nur nach den Elementen der ersten Zeile oder Spalte entwickelt zu werden braucht und auf die verbleibenden Determinanten 3. Ordnung die Regel von SARRUS¹²²⁾ angewandt werden kann. Die charakteristische Gleichung hat die Form

$$(54''') \quad \text{Det}(\mathcal{R} - \lambda \mathcal{E}) = \lambda^4 - 4\lambda^3 + 4,504024\lambda^2 - 1,675223\lambda + 0,121712 = 0.$$

Wir vergleichen dieses Ergebnis mit (54') und stellen fest, daß die Vorzeichen der Koeffizienten

$c_1 = - 4$
$c_2 = + 4,504024$
$c_3 = - 1,675223$
$c_4 = + 0,121712$

den Anforderungen entsprechen. Weiterhin sind die Bedingungen ¹²³⁾

$$c_1 = \sum_{i=1}^4 |x_{1i}| = |4|$$

und

$$c_4 = \text{Det } \mathcal{R} = 0,121712$$

erfüllt. Zur Lösung einer algebraischen Gleichung 4. Grades stehen verschiedene Möglichkeiten zur Verfügung. Wir wählen einen Weg, der die Aufspaltung in zwei quadratische Gleichungen

$$(\lambda^2 + f_1\lambda + g_1)(\lambda^2 + f_2\lambda + g_2) = 0$$

vorsieht ¹²⁴⁾. Hierbei können durch Koeffizientenvergleich die Beziehungen

$$\begin{aligned} f_1 + f_2 &= c_1 \\ f_1f_2 + g_1 + g_2 &= c_2 \\ f_1g_2 + f_2g_1 &= c_3 \\ g_1g_2 &= c_4 \end{aligned} \quad (141)$$

gewonnen werden. Man geht nun so vor, daß g_1 oder g_2 innerhalb des Bereichs

$$(141') \quad 0 \leq g \leq +\sqrt{|c_4|}$$

mit

$$+\sqrt{c_4} \sim 0,3489$$

frei gewählt werden und probiert so lange, bis alle Beziehungen (141) erfüllt werden. Wir erhalten schließlich mit

123) Vgl. Fußnote 41 auf Seite 31.

124) Dieses Verfahren erscheint deshalb vorteilhaft, weil es das Aufstellen einer kubischen Resolvente vermeidet, deren Lösungen entweder graphisch oder mit kubischen Wurzeln (CARDANISCHE Formeln) gefunden werden müssen.

Vgl. ZURMÜHL, R. (1957), a.a.O., Seite 59-62.

$$\begin{aligned} f_1 &= -1,0010792 \\ f_2 &= -2,9989208 \\ g_1 &= +0,0859605 \\ g_2 &= +1,4159061 \end{aligned}$$

die Gleichungen

$$(\lambda^2 - 1,0010792\lambda + 0,0859605)(\lambda^2 - 2,9989208\lambda + 1,4159061) = 0$$

und daraus die Eigenwerte der Korrelationsmatrix

$$\begin{aligned} \lambda_1 &= 2,4118613 \\ \lambda_2 &= 0,9062235 \\ \lambda_3 &= 0,5870595 \\ \lambda_4 &= 0,0948557 \end{aligned}$$

Diese sind, wie erwartet, alle reell und positiv. Sie erfüllen die charakteristische Gleichung und genügen den aus den Wurzelsätzen von VIETA abgeleiteten Bedingungen

$$(106') \quad \text{Sp } \mathcal{R} = \sum_{i=1}^4 r_{ii} = \sum_{k=1}^4 \lambda_k = 4$$

sowie

$$(142) \quad \prod_{k=1}^4 \lambda_k = \text{Det } \mathcal{R} = 0,121712.$$

Das erste Ergebnis der Eigenwertaufgabe ist die Feststellung, daß die Korrelationsmatrix den vollen Rang 4 hat und demzufolge für die Analyse ($p=n$) gilt. Allerdings ist λ_4 klein genug, um die Hauptkomponente z_4 vielleicht vernachlässigen zu können. Darüber gibt die folgende Übersicht, die aus (39) und (105) folgt¹²⁵⁾, Aufschluß.

Tabelle 3: Anteilswerte der Hauptkomponenten an der Gesamtvariation der Ausgangsvariablen

Eigenwert	Varianzanteile (in v.H.)	
	je Eigenwert	kumuliert
(1)	(2)	(3)
1	60,29653	60,29653
2	22,65558	82,95211
3	14,67648	97,62859
4	2,37139	99,99998
Summe	99,99998	

125) Vgl. die entsprechenden Ausführungen auf Seite 52.

Wie man sieht, werden bereits durch die ersten drei Hauptkomponenten 97,6 % der Varianzen der Ausgangsvariablen erfaßt; das System der x_1, x_2, x_3, x_4 kann daher nahezu ohne Informationsverlust durch das der z_1, z_2, z_3 dargestellt werden.

212 Eigenvektoren und Iterationsprozeß

Die Berechnung der Eigenvektoren geschieht in der in Abschnitt 201 geschilderten Weise. Wir haben vier lineare Gleichungssysteme mit je vier Unbekannten $a_{1k}^0, a_{2k}^0, a_{3k}^0$ und a_{4k}^0 ($k=1,2,3,4$) zu lösen. Jedes System entsteht durch Einsetzen des Eigenwertes λ_h ($h=1,2,3,4$) in die Beziehungen (53'), wobei in allen Fällen die erste Koordinate gleich 1 gesetzt wird.

Tabelle 4: Eigenvektoren der Korrelationsmatrix

n \ p	Eigenvektoren			
	α_1^0	α_2^0	α_3^0	α_4^0
(1)	(2)	(3)	(4)	(5)
1	1	1	1	1
2	-0,834794	-0,797152	8,334672	0,496489
3	-0,430004	4,200065	1,080086	-0,117705
4	0,902944	0,155753	7,112397	-0,704626

Diese Vektoren werden nach (132) normiert und bilden die im Abschnitt 130 definierte Matrix

$$A = \begin{bmatrix} 0,945645 & 0,216688 & 0,069304 & 0,232362 \\ -0,787419 & -0,172735 & 0,577624 & 0,115365 \\ -0,406631 & 0,910107 & 0,074853 & -0,027350 \\ 0,853865 & 0,033750 & 0,492915 & -0,163729 \end{bmatrix},$$

mit deren Hilfe wir das System (11) in der Form

$$(143) \quad \begin{aligned} x_1 &= 0,945645z_1 + 0,216688z_2 + 0,069304z_3 + 0,232362z_4 \\ x_2 &= -0,787419z_1 - 0,172735z_2 + 0,577624z_3 + 0,115365z_4 \\ x_3 &= -0,406631z_1 + 0,910107z_2 + 0,074853z_3 - 0,027350z_4 \\ x_4 &= 0,853865z_1 + 0,033750z_2 + 0,492915z_3 - 0,163729z_4 \end{aligned}$$

schreiben können.

Obwohl bei der hier gegebenen Reihenzahl der Korrelationsmatrix auf alle Fälle das direkte Verfahren der Eigenwertbestimmung vorzuziehen ist, wurde für λ_1 und λ_2 der einfache Iterationsprozeß durchgeführt. Ausgehend von einem Vektor

$$b_0 = \{1 \quad 1 \quad 1 \quad 1\}$$

führte dies für λ_1 nach 17 Iterationen auf

$$b_{17} = \{1 \quad -0,834793 \quad -0,430002 \quad 0,902943\}$$

mit einem bei weiteren Iterationen konstanten Verhältnis

$$q_{i,17} = \frac{b_{i,17}}{b_{i,16}} = 2,41186 \quad (i=1,2,3,4).$$

Damit konnten Eigenwert λ_1 und Eigenvektor a_1^0 der Tabelle 4 verhältnismäßig rasch gut angenähert werden. Für λ_2 wurde von einem zu b_{17} orthogonalen Vektor¹²⁶⁾

$$b'_0 = \{1 \quad 1 \quad 1 \quad 0,29325\}$$

ausgegangen. Hier ergab sich erst auf fortgeschrittener Iterationsstufe bei wiederholter "Reinigung" von b_y ein einigermaßen stabiler Vektor

$$b'_{74} = \{0,239 \quad -0,189 \quad 1 \quad 0,038\},$$

der in der Form

$$b'_{74} = \{1 \quad -0,791 \quad 4,184 \quad 0,159\}$$

mit a_2^0 der Tabelle 4 nur in grober Annäherung übereinstimmt. Das Verhältnis (134) nimmt die Form

$$q'_{i,74} = \frac{b'_{i,74}}{b'_{i,73}} \sim 0,906$$

an. Die Ursache dieser ungenügenden Konvergenz ist wohl darin zu suchen, daß das schon ziemlich kleine Verhältnis

126) Zur Reinigung des Iterationsvektors bei höheren Eigenwerten (Verfahren von KOCH) siehe ZURMÜHL, R. (1958), a.a.O., Seite 292-294.

$$\lambda_1/\lambda_2 = 2,661440$$

bei dem zweiten Eigenvektor auf

$$\lambda_2/\lambda_3 = 1,543665$$

fällt. Vermutlich wäre die Konvergenz von nun an wegen

$$\lambda_3/\lambda_4 = 6,188960$$

wieder wesentlich besser geworden. Wir haben diesen Weg hier nicht weiter verfolgt. Man kann annehmen, daß sich durch Iteration mit den Matrizen $R_{.1}^*$, $R_{.2}^*$ und $R_{.3}^*$ in kürzerer Zeit bessere Ergebnisse erzielen ließen.

213 Die Hauptkomponenten; Interpretation der Ergebnisse

Da die Matrix A jetzt bekannt ist, können nach (111) aus

$$(144) \quad z_k = \frac{1}{\lambda_k} a'_k x \quad (k=1,2,3,4),$$

nämlich dem System

$$(144') \quad \begin{aligned} z_1 &= 0,392081x_1 - 0,326478x_2 - 0,168596x_3 + 0,354027x_4 \\ z_2 &= 0,239111x_1 - 0,190610x_2 + 1,004285x_3 + 0,037242x_4 \\ z_3 &= 0,118053x_1 + 0,983928x_2 + 0,127505x_3 + 0,839634x_4 \\ z_4 &= 2,449637x_1 + 1,216216x_2 - 0,288385x_3 - 1,726085x_4 \end{aligned}$$

die Hauptkomponenten berechnet werden. Als Zeilen angeordnet, ergeben sie die Matrix ¹²⁷⁾

- 127) Durch schwer erfaßbare Einflüsse (z.B. Rundungsfehler) erreichen die Hauptkomponenten z_1, z_2, z_3, z_4 nicht ganz die vorgeschriebene Länge $\sqrt{6} = 2,449490$. Die einzelnen Längen sind:
- | | | | |
|-------|---|----------|------------|
| z_1 | : | 2,486491 | (+1,51 %) |
| z_2 | : | 2,443027 | (-0,30 %) |
| z_3 | : | 2,306814 | (-5,83 %) |
| z_4 | : | 2,419876 | (-1,21 %). |

Infolgedessen sind die Hauptkomponenten auch nicht streng orthogonal. Der Grund dieser Erscheinung dürfte unter anderem auch darin liegen, daß R mit $\lambda_4 = 0,094$

"nahezu singular" ist, also "fast" den Rang 3 statt 4 hat, wodurch eine gewisse Inkonsistenz in die Gleichungen (53') hineingerät. Bei den Eigenvektoren der Tabelle 4 konnte dies durch iterative Nachbehandlung ausgeglichen werden, aber bei den Hauptkomponenten muß man in diesem Falle die relativ geringe Fehlerspanne in Kauf nehmen.

$$(109') \quad \mathbf{z} = (\mathbf{a}'\mathbf{a})^{-1}\mathbf{a}'\mathbf{x} =$$

-1,281124	-1,141941	0,876479	-0,425017	1,270622	0,820972
-0,420191	-0,662866	0,501420	1,882258	-0,040379	-1,247590
0,909452	-0,104407	1,454902	-1,110203	0,183801	-1,048958
-1,109766	0,887545	1,297029	-0,033261	-1,457567	-0,169088

Damit sind wir am Ziel unserer Analyse und können im Sinne der Aufgabenstellung (11) die Ausgangsvariablen je nach Genauigkeitsbedürfnis durch die ersten zwei, drei oder vier Hauptkomponenten ersetzen. Es empfiehlt sich jedoch, die Anteilswerte der z_k nicht nur an der Gesamtvariation, sondern auch an den Varianzen der einzelnen Variablen zu ermitteln und damit einen besseren Einblick in die Struktur der Hauptkomponenten zu gewinnen. Das geschieht nach (38) sehr einfach durch das Quadrieren der Elemente von $\mathbf{a}$ und liefert die folgende Übersicht.

Tabelle 5: Anteilswerte der Hauptkomponenten an den Varianzen der Ausgangsvariablen

Hauptkomponente	Varianzanteile je Hauptkomponente (in v.H.)			
	x_1	x_2	x_3	x_4
(1)	(2)	(3)	(4)	(5)
z_1	89,4245	62,3182	16,5349	72,9085
z_2	4,6954	2,9837	82,8294	0,1139
z_3	0,4803	33,3649	0,5603	24,2965
z_4	5,3992	1,3309	0,0748	2,6807
Summe	99,9994	99,9977	99,9994	99,9996

Durch fortlaufende Aufsummierung der Spalten erhalten wir

Tabelle 6: Kumulierte Anteilswerte der Hauptkomponenten an den Varianzen der Ausgangsvariablen

Hauptkomponente	Kumulierte Varianzanteile (in v.H.)			
	x_1	x_2	x_3	x_4
(1)	(2)	(3)	(4)	(5)
z_1	89,4245	62,3182	16,5349	72,9085
z_2	94,1199	65,3019	99,3643	73,0224
z_3	94,6002	98,6668	99,9246	97,3189
z_4	99,9994	99,9977	99,9994	99,9996

Diese Aufschlüsselungen zeigen erneut, wie stark die Variation auch der einzelnen Variablen bereits bei Erreichen der dritten Hauptkomponenten ausgeschöpft worden ist. Sie zeigt aber weitergehend, daß die Varianzanteile der z_k an den einzelnen x_i sehr unterschiedlich ausfallen und sich erst bei z_3 einem einheitlichen Niveau nähern. Dieser Umstand ist hier für uns nicht sehr wesentlich, er stellt aber - wie wir im Abschnitt 310 sehen werden - einen der Hauptgesichtspunkte in der Faktoranalyse dar. Dort nämlich versucht man, die x_i durch die z_k zu "erklären", etwa in dem Sinne, daß die Variable x_i durch einen zugrunde liegenden "Faktor" z_i mit 89,4 % "erklärt" wird, für praktische Zwecke also mit ihm identisch ist.

Zum Abschluß sei nun noch mit Hilfe der Beziehungen

$$(98'') \quad r_{ih.4} = r_{ih} - \sum_{k=1}^4 a_{ik} a_{hk} \quad (i, h=1, 2, 3, 4)$$

die Matrix

$$(99') \quad R_{.4}^* = \tilde{\alpha}_{.4} \tilde{\alpha}_{.4}'$$

$$= \begin{bmatrix} 0,000006 & 0,000012 & -0,000002 & 0,000007 \\ 0,000012 & 0,000023 & 0,000002 & 0,000007 \\ -0,000002 & 0,000002 & 0,000006 & -0,000002 \\ 0,000007 & 0,000007 & -0,000002 & 0,000004 \end{bmatrix}$$

aufgestellt. Sie enthält die Residuen der Korrelationskoeffizienten nach Abzug der Auswirkung der vier Hauptkomponenten und zeigt, daß die Umstellung auf diese in nahezu vollständigem Maße durchgeführt werden kann. Im übrigen sei hier abermals darauf verwiesen ¹²⁸⁾, daß mit Hilfe der Koeffizienten a_{ik} die Korrelationsmatrix "reproduzierbar" ist.

Die Bedeutung der einzelnen Stufen der durchgeführten Rechnung für die zu Beginn des Abschnitts angedeutete Anwendungsmöglichkeit auf vier wirtschaftsstatistische Merkmale sei wie folgt skizziert. Zu den vier Ausgangsvariablen wurden vier Hauptkomponenten bestimmt, von denen die ersten

128) Vgl. die entsprechende Bemerkung auf Seite 26.

beiden zusammen etwa 83 %, die ersten drei über 97 % der Gesamtvariation wiedergeben (Tabelle 3). Es ist daher möglich, in bestimmten ökonomischen Modellbildungen anstelle von vier Variablen nur zwei oder drei Hauptkomponenten zu verwenden, die für jeden Landkreis aus (144') berechnet werden können. Eine Verringerung der Dimension in diesem Maße ist natürlich trivial, ihr Nutzen erst bei wesentlich größeren als der hier aus rechentechnischen Gründen klein gehaltenen Ausgangsgesamtheit voll erkennbar. Die Beziehungen zwischen Variablen und Hauptkomponenten lassen sich dahingehend erweitern, daß die Varianzanteile der letzteren an den einzelnen Merkmalen Untersuchungsgegenstand werden (Tabellen 5 und 6). So hat die erste Hauptkomponente an der Variation aller vier Merkmale zwar einen Anteil von 60 %, von der Varianz des ersten Merkmals erfaßt sie dagegen bereits 89 % und von der des dritten Merkmals nur knapp 17 % (Tabelle 5, Spalten 2 und 4). Diese Unterschiede sind dann von Interesse, wenn der Versuch unternommen wird, den Hauptkomponenten auf der Grundlage der sachlichen Beschaffenheit der Ausgangsmerkmale eine "reale" Bedeutung beizulegen, wozu übrigens auch die gegenseitigen Abhängigkeitsbeziehungen der Systeme (143) und (144') herangezogen werden können ¹²⁹⁾.

129) Diese Ausführungen sollen das Zahlenbeispiel durch einige Hinweise auf die praktische Anwendung der Komponentenanalyse vervollständigen, eine ausführliche Diskussion der Anwendungsfragen bleibt jedoch wegen der dazu notwendigen Ergänzungen des Abschnitts 3 dem Abschnitt 4, Seite 122 ff., vorbehalten. Dieses Vorgehen empfiehlt sich auch aus den Gründen, die wir bereits auf Seite 74 einer allzu konkreten Deutung des Zahlenbeispiels entgegen gesetzt haben.

3 ENTWICKLUNG DER KOMPONENTENANALYSE UND ABGRENZUNG
 GENÜßER DER FAKTORANALYSE

30 EINIGE BEITRÄGE ZUR ENTWICKLUNG DER KOMPONENTEN-
 ANALYSE

300 Vorbemerkung

Die Komponentenanalyse wurde im Abschnitt 1 formal entwickelt und im Abschnitt 2 unter dem Gesichtspunkt ihrer numerischen Anwendung behandelt. Dabei konnten wir uns stets auf eine in sich geschlossene Ableitung des Verfahrens stützen, die es erlaubte, auf die Berücksichtigung der Entwicklung der Komponentenanalyse in der Literatur im wesentlichen zu verzichten. Ausnahmen davon bildeten lediglich die allgemeine Bezugnahme auf HOTELLING sowie vereinzelte Hinweise auf weitere Autoren, namentlich bei der Behandlung der Rechentechnik ¹³⁰⁾.

War dieses Vorgehen durch die sich selbst tragende Entwicklung der Methode bisher gerechtfertigt, so kann der Gesamtaufbau der Arbeit einer wenigstens als Überblick gestalteten Eingliederung in die unterschiedlichen Verfahrensweisen der Komponentenanalyse nicht entraten. Dies muß besonders deshalb unerläßlich erscheinen, weil die für den Abschnitt 4 vorgesehenen wirtschafts- und sozialwissenschaftlichen Anwendungen in ihren methodischen Voraussetzungen voneinander abweichen und daher eine vorbereitende Orientierung erfordern. Hinzu kommt, daß die durchaus noch ungeklärte Abgrenzung der Komponentenanalyse gegenüber der Faktoranalyse - eigentlich die Frage, ob überhaupt eine "selbständige" Komponentenanalyse existiert - auch hinsichtlich der Anwendungen einer Erörterung bedarf.

130) Vgl. insbesondere Seite 1 und die Quellenangaben zum Abschnitt 20, Seite 61-73. Von HOTELLING haben wir gewissermaßen den "Rahmen" unserer Ausführungen entlehnt, das HOTELLINGSCHES Verfahren selbst stimmt mit dem unseren in den Voraussetzungen und in der Durchführung nur teilweise überein.

Wir gehen nun so vor, daß der Reihe nach gewisse Formen der Komponentenganalyse geschildert und mit unserem eigenen Vorgehen verglichen werden, wobei es erlaubt sein möge, die Darstellung auf die Ergebnisse zu konzentrieren und auf deren Herleitung im einzelnen zu verzichten. Die Berührung mit der Faktoranalyse, bisher wegen der Gefahr des Hinübergleitens in rein psychologische Sachverhalte hinausgeschoben, soll dazu den vorläufigen Abschluß bilden.

301 Die "Geraden bester Anpassung" von KARL PEARSON und ihre Deutung durch KENDALL

Eine beträchtliche Zahl der in der statistischen Theorie der Gegenwart gebräuchlichen Methoden hat ihre Wurzel in den Arbeiten, die kurz vor der Jahrhundertwende von GALTON und K. PEARSON auf dem Gebiete der biologischen Statistik - der Biometrie im heutigen Sinne - durchgeführt wurden. Gilt dies ganz allgemein für die Begriffe der Regression und der Korrelation, so kann speziell auch der Ursprung der Komponentenganalyse in der bereits zu Beginn erwähnten Veröffentlichung ¹³¹⁾ von PEARSON gesehen werden, die als erster mathematischer Ausdruck für Probleme der Untersuchung einer Gesamtheit von Variablen hinsichtlich zugrunde liegender Ursachenkomplexe, durch die die Variationen in den Variablen bewirkt werden, zu gelten hat ¹³²⁾.

131) Vgl. Fußnote 2 auf Seite 1.

132) Es interessieren hier nur die ersten methodischen Ansätze. Zur Frage des sachlichen Hintergrundes siehe PEARSON, K. (1896) - "Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia", Roy. Soc. Lond. Phil. Trans. A. 187, Seite 253-318, insbesondere Seite 301, sowie EDGEWORTH, F. Y. (1896) - "Supplementary notes on statistics", J. Roy. Stat. Soc. 59, Seite 529-539, insbesondere Seite 539.

Eine eingehende Diskussion mit weiteren Literaturhinweisen ist enthalten in BURT, C. (1949) - "Alternative methods of factor analysis and their relations to PEARSON'S method of 'principal axes'", Brit. J. Psychol. Stat. Sect. 2, Seite 98-121.

PEARSON geht davon aus, daß die Untersuchung eines Komplexes von untereinander korrelierten Variablen damit beginnen müsse, die Variablen durch dieselbe Anzahl von unkorrelierten Komponenten - die er "ideal index characters" nennt - zu ersetzen. Da zu dieser Aufgabe, wie auch unsere Ableitung gezeigt hat, keine eindeutige Lösung existiert, soll die Variation in der Ausgangsgesamtheit von Stufe zu Stufe durch die Komponenten verringert werden. Dem entspricht in geometrischer Analogie die Zielsetzung, eine Schar von Punkten in einem Raum höherer Dimension durch eine "Gerade bester Anpassung" (straight line of best fit) darzustellen, die dadurch ausgezeichnet ist, daß die Summe der Entfernungsquadrate ¹³³⁾ zwischen den Punkten und der Geraden zu einem Minimum wird ¹³⁴⁾. Man kann nun zeigen, daß eine Verallgemeinerung dieses Problems auf ebenso viele dieser Geraden wie es Ausgangsvariable gibt zu der von uns besprochenen

133) Hiermit sind die Lote von den Punkten auf die Gerade gemeint. PEARSON (1901, a.a.O., Seite 559-560) schreibt: "In many physical, statistical, and biological investigations it is desirable to represent a system of points in plane, three, or higher dimensional space by the 'best fitting' straight line or plane. ... Of course the term 'best fit' is really arbitrary; but a good fit will clearly be obtained if we make the sum of the squares of the perpendiculars from the system of points upon the line or plane a minimum".

134) Weitere Ausführungen dazu finden sich bei WHIPPLE, W.H. (1926) - "On the best linear relation connecting three variables", PhilMag. 7th Ser. 1, Seite 378-384.
RHODES, E.C. (1927) - "On lines and planes of closest fit", PhilMag. 7th Ser. 3, Seite 357-364.
UVEN, M.J. VAN (1930) - "Adjustment of N points of n-dimensional space to the best linear (n-1)-dimensional space", Proc. Amsterdam 33, Seiten 143-157 und 307-326.
Es sei noch angemerkt, daß die hier geschilderte Gerade (im zweidimensionalen Fall) nicht mit der üblichen Regressionsgeraden, wie sie mit Hilfe der Methode der kleinsten Quadrate ermittelt wird, übereinstimmt, weil in unserer Betrachtung die bei der Regressionsanalyse anzutreffenden Abhängigkeitsbeziehungen zwischen den Variablen nicht unterstellt werden. Es handelt sich vielmehr um die sogenannte "orthogonale Regression", soweit man diesen Begriff gelten lassen will. Vgl. dazu CRAMÉR, H. (1946), a.a.O., Seite 275 und Seite 309-310 sowie KENDALL-BUCKLAND (1957), a.a.O., Seite 208.

Hauptachsentransformation führt; dabei bilden die "Geraden bester Anpassung" das System der Hauptachsen, auf die die Koordinaten der Punkte bezogen werden. Der somit angedeutete Lösungsweg von PEARSON ist in neuerer Zeit von KENDALL aufgegriffen worden ¹³⁵⁾. Wir können uns daher auf einen Vergleich unserer Methode mit den Ausführungen von KENDALL beschränken.

Die Ausgangsvoraussetzungen stimmen mit denen der Abschnitte 110 bis 112 überein, einschließlich der Normierung der Variablen X_i ($i=1,2,\dots,n$) nach (4). Die Komponenten werden jedoch lediglich als "zentriert" mit dem Mittelwert 0, sonst aber beliebigen Varianzen unterstellt. Gegenüber (6) wollen wir sie daher mit z'_k ($k=1,2,\dots,p$) bezeichnen. Die Werte der Ausgangsgesamtheit lassen sich wieder in einer Matrix $\mathcal{A}$ anordnen ¹³⁶⁾. Während wir bisher in dieser Matrix die Zeilen als n Vektoren im N -dimensionalen Raum definiert haben, betrachten wir jetzt die Spalten als die Koordinaten von N Punkten im n -dimensionalen euklidischen Raum R_n ¹³⁷⁾. Für den laufenden Punkt P ($Y_1; Y_2; \dots; Y_n$) einer Geraden im R_n gilt die Beziehung

135) KENDALL, M.G. (1950) - "Factor analysis as a statistical technique" (Teil 1 von KENDALL, M.G.; BABINGTON SMITH, B. - "Factor analysis"), J. Roy. Stat. Soc. B. 12, Seite 60-73.
- - - Ders. (1957) - A Course in Multivariate Analysis (Griffin's Statistical Monographs & Courses, No. 2). London: Charles Griffin & Co., Seite 10-36.
Eine weitere Darstellung der von PEARSON ausgehenden Deutung des Komponentenproblems stammt von RAO, C.R. (1949) - "Representation of 'p' dimensional data in lower dimensions" ("Anthropometric survey of the United Provinces, 1941: A statistical study", Appendix 4), Sankhyā 9, Seite 248-251.

136) Vgl. Seite 16.

137) Natürlich können wir auch in diesem Fall die n Punkte als Endpunkte von N Radiusvektoren im V_n ansehen und damit die Vektoralgebra erneut heranziehen. Wir folgen jedoch KENDALL und beziehen uns auf einen lineären Punkttraum mit fester Orthonormalbasis, zumal die Deutung der Korrelationen als Winkel zwischen Vektoren nunmehr hinfällig ist.

$$(145) \quad \frac{Y_1 - M_1}{l_1} = \frac{Y_2 - M_2}{l_2} = \dots = \frac{Y_n - M_n}{l_n},$$

wobei die M_i ($i=1,2,\dots,n$) die Koordinaten des Schwerpunktes und die l_i ($i=1,2,\dots,n$) die Richtungskosinus sind, für die bei dem hier gegebenen rechtwinkligen Koordinatensystem

$$(146) \quad \sum_{i=1}^n l_i^2 = 1$$

gilt ¹³⁸⁾. Bezeichnen wir mit ND die Summe der senkrechten Entfernungsquadrate der vorgegebenen N Punkte auf die Gerade (145), so ist ¹³⁹⁾

$$(147) \quad ND = \sum_{j=1}^N \left[\sum_{i=1}^n (x_{ij} - M_i)^2 - \left\{ \sum_{i=1}^n l_i (x_{ij} - M_i) \right\}^2 \right].$$

In (145) können die Größen M_i und l_i , letztere unter Beachtung von (146), frei gewählt werden. Die partielle Differentiation von (147) nach allen M_i führt darauf, daß die Gerade durch den Schwerpunkt des Punktesystems läuft, der aber wegen (5) mit dem Ursprung der Achsen zusammenfällt. Wir können deshalb alle $M_i=0$ setzen und erhalten

$$(148) \quad ND = \sum_{j=1}^N \left[\sum_{i=1}^n x_{ij}^2 - \left(\sum_{i=1}^n l_i x_{ij} \right)^2 \right],$$

daraus wegen (13) weiterhin

$$(149) \quad D = n - \frac{1}{N} \sum_{j=1}^N \left(\sum_{i=1}^n l_i x_{ij} \right)^2.$$

Wird nun für D unter den Nebenbedingungen (146) ein Extremwert gesucht, so liefert die Differentiation n Gleichungen der Form

¹³⁸⁾ Ein Richtungskosinus ist der Kosinus des Winkels, den ein aus dem Ursprung kommender und mit der Geraden (145) parallel verlaufender Strahl mit der positiven Richtung einer der Koordinatenachsen bildet. Vgl. MANGOLDT-KNOPP (1948), a.a.O., 1, Seite 268-271.

¹³⁹⁾ Zur folgenden Ableitung siehe PEARSON, K. (1901), a.a.O., Seite 563-564; KENDALL, M. G. (1950), a.a.O., Seite 63, Gleichung (10) bis (13); Ders. (1957), a.a.O., Seite 14, Gleichung (2.8) bis (2.15).

$$(150) \quad \frac{1}{N} \sum_{j=1}^N x_{ij} \sum_{h=1}^n l_h x_{hj} - \lambda l_i = 0 \quad (i=1,2,\dots,n),$$

die in der bekannteren Schreibweise

$$(150') \quad \sum_{h=1}^n r_{ih} l_h - \lambda l_i = 0 \quad (i=1,2,\dots,n)$$

dem System (53') entsprechen ¹⁴⁰⁾. Da es sich hier um die Bestimmung der ersten Hauptkomponenten z_1' handelt, wären zur Angleichung an (53') statt l_i oder l_h die Symbole l_{i1} beziehungsweise l_{h1} zu verwenden. An die Stelle der Koeffizienten a_{ik} treten im allgemeinen Falle der Hauptkomponenten z_k' ($k=1,2,\dots,p$) die Richtungskosinus l_{ik} . Damit stimmt - die geometrische Analogie ausgenommen - die an (150') anzuschließende Diskussion des Eigenwertproblems der Korrelationsmatrix mit unseren Ausführungen im Abschnitt 15 völlig überein.

Wird (150') mit l_i multipliziert und über i summiert, so erhält man unter Beachtung von (146) den Ausdruck

$$(151) \quad \lambda = \sum_{i,h=1}^n r_{ih} l_i l_h = \frac{1}{N} \sum_{j=1}^N \left(\sum_{i=1}^n l_i x_{ij} \right)^2.$$

Daher folgt aus (149)

$$(149') \quad D = n - \lambda.$$

Das heißt aber, daß bei gegebenem n die Größe D mit steigendem λ fällt - mit anderen Worten, wir erhalten die "beste" Gerade (145) mit einem Minimum für (147) dann, wenn unter den Eigenwerten der Korrelationsmatrix der größte Wert λ_1 - im Sinne von (64) - ausgewählt wird. Ist λ_1 bekannt, so können aus (150') die n Größen l_{i1} berechnet und nach

$$(152) \quad z_1' = \sum_{i=1}^n l_{i1} x_i$$

für die Bestimmung der ersten Hauptkomponenten verwendet werden.

Hinsichtlich der auf z'_1 folgenden Hauptkomponenten gelten auch hier in entsprechender Übertragung die Überlegungen des Abschnitts 16, so daß wir allgemein

$$(153) \quad z'_k = \sum_{i=1}^n l_{ik} x_i \quad (k=1, 2, \dots, p)$$

schreiben können. Hierbei macht es wieder keinen Unterschied, ob die x_i linear abhängig ($p < n$) oder unabhängig ($p = n$) sind ¹⁴¹⁾.

Wir wenden uns abschließend den Varianzen der z'_k zu. Aus (153) folgt sofort, daß ¹⁴²⁾

$$(154) \quad \text{Var } z'_k = \frac{1}{N} \sum_{j=1}^N \left(\sum_{i=1}^n l_{ik} x_{ij} \right)^2 \quad (k=1, 2, \dots, p),$$

im Vergleich zu (151) ist daher

$$(151'') \quad \text{Var } z'_k = \lambda_k.$$

Damit ist wegen (106) die Ausschöpfung der Gesamtvariation sichergestellt. Überdies lassen sich durch

$$(155) \quad z_k = \frac{1}{\sqrt{\lambda_k}} z'_k \quad (k=1, 2, \dots, p)$$

die Hauptkomponenten (153) in die unserer vorausgegangenen Ableitung überführen ¹⁴³⁾. Wenn wir, um KENDALL zu folgen, den in diesem Abschnitt betrachteten R_n als " n_1 -Raum", den zu Beginn definierten T_p - wobei wir der Einfachheit halber ($p=n$) annehmen - als " n_2 -Raum" bezeichnen, so sind nur die Hauptkomponenten, nach (111) oder (153) berechnet, in beiden

141) Vgl. die Ausführungen zum Rang von $\mathcal{H}$ und $\mathcal{R}$ in Abschnitt 131, Seite 19-20. Diese gelten wegen der Gleichheit von Zeilen- und Spaltenrang auch in der jetzt vorliegenden Betrachtungsweise.

142) (154) kommt dadurch zustande, daß wegen (5)

$$z'_k = \frac{1}{N} \sum_{j=1}^N \sum_{i=1}^n l_{ik} x_{ij} = 0 \text{ gelten muß.}$$

143) Vgl. dazu auch (111'') auf Seite 53 sowie KENDALL, M.G. (1957), a.a.O., Seite 25-26.

Raumvorstellungen orthogonal¹⁴⁴⁾. Insbesondere sind Komponenten, die nach gewissen Methoden der Faktoranalyse bestimmt werden, wohl unkorreliert, also orthogonal im " n_2 -Raum", nicht aber gleichzeitig orthogonal im " n_1 -Raum".

302 Die Komponentenanalyse nach HOTELLING, GIRSHICK und
VAN DEN BOGAARD-KAPTEIN

Wenn das ursprünglich von HOTELLING vorgeschlagene Verfahren kurz betrachtet werden soll, so eröffnet sich neben der geschilderten Entwicklungslinie von PEARSON zu KENDALL und unserer eigenen Ableitung eine dritte Variante der Komponentenanalyse, jedoch ist der Unterschied zu gering, um eingehende Ausführungen zu rechtfertigen. HOTELLING benutzt ebenfalls die Raumvorstellung, die wir im vorigen Abschnitt als " n_1 -Raum" bezeichnet haben, nimmt aber die Hauptkomponenten als normiert nach (6) an, so daß sie im Endergebnis in der Form (111'') erscheinen¹⁴⁵⁾. Andererseits aber ermöglicht die Deutung der Ausgangsmatrix $\mathcal{A}$ im Sinne von PEARSON die uns verschlossen gebliebene Anwendung der

144) KENDALL (1950, a.a.O., Seite 66) verwendet p anstelle von n sowie n statt N und schreibt: "This p -space is quite different from the space previously considered. It may be as well to give it a distinguishing name and I shall accordingly call it the p_2 -space, whereas the space of n points will be called the p_1 -space. There is a duality between them in the broad sense (arising from the fact that we may scan the observation matrix across or down), but there is no geometrical correspondence in the customary sense. In particular a set of axes which orthogonal in one is not necessarily orthogonal so in the other. ... Much of the literature is concerned with 'orthogonal components'. This nearly always means components which are orthogonal in the p_2 -space, i. e. are uncorrelated. Only the principal components ... are orthogonal in both the p_1 - and the p_2 -spaces".

145) HOTELLING, H. (1933), a.a.O., Gleichung (21) auf Seite 426.

Hauptachsentransformation ¹⁴⁶⁾ } zumal die z_k als normal
verteilt angenommen werden ¹⁴⁷⁾;

"If $z_1, z_2, \dots, z_n$ be taken as rectangular coordinates in n dimensions, each point represents a possible individual. If, as we assume, the population is normally distributed, the loci of uniform density are concentric, similar, and similarly placed ellipsoids. The method of principal components ... is equivalent to choosing a set of coordinate axes coinciding with the principal axes of these ellipsoids".

und ¹⁴⁸⁾

"Geometrically, the foregoing procedure corresponds to rotating the rectangular axes of $z_1, z_2, \dots, z_n$ so that the new coordinate axes lie along the principal axes of the ellipsoids of uniform density. The squares of the lengths of the principal axes of one of these ellipsoids are proportional to the k 's".

Mit Ausnahme einer kurzen Bemerkung in der Einleitung geht HOTELLING stets von unserem Fall ($p=n$) mit einer nichtsingulären Verteilung der x_i aus. Diese Lücke wurde wenige Jahre später von GIRSHICK ¹⁴⁹⁾ durch den Nachweis geschlossen, daß die Methode der Komponentenanalyse nicht darauf angewiesen ist, zu n Ausgangsvariablen in jedem Falle auch n Hauptkomponenten bestimmen zu müssen. Dies entspricht der von uns vorgesehenen Möglichkeit ($p \leq n$), wozu wir insbesondere die Ableitung (44') bis (51) ¹⁵⁰⁾ von GIRSHICK übernommen haben ¹⁵¹⁾. GIRSHICK bringt außerdem drei weitere methodische Ergänzungen, deren Ergebnisse wir nachstehend kurz schildern wollen.

Es sei angenommen, daß die normierten Ausgangsvariablen x_i aus zwei Bestandteilen

146) Vgl. Seite 58-60.

147) HOTELLING, H. (1933), a.a.O., Seite 421.

148) - - - Ders. (1933), a.a.O., Seite 427. HOTELLING versteht die normierten Ausgangsvariablen mit dem Symbol z , die Werte k sind die Eigenwerte der Korrelationsmatrix. Zum letzten Satz siehe (123) auf Seite 57.

149) GIRSHICK, M.A. (1936) - "Principal components", J. Am. Stat. Ass. 31, Seite 519-528.

150) Vgl. Seite 28-29.

151) GIRSHICK, M.A. (1936), a.a.O., Gleichungen (5) bis (10) auf Seite 522.

$$(156) \quad x_i = \alpha_i + \beta_i \quad (i=1,2,\dots,n)$$

bestehen. Dabei sollen die α_i die "wahren" Messungen der Variablen x_i sein, die durch einen zufälligen Fehler β_i , für den die Annahmen

$$(157) \quad \bar{\beta}_i = \frac{1}{N} \sum_{j=1}^N \beta_{ij} = 0 \quad (i=1,2,\dots,n),$$

$$(158) \quad \text{Var } \beta_i = \text{Var } \beta_h \quad (i,h=1,2,\dots,n)^{152)}$$

und

$$(159) \quad r_{ih} = \frac{\sum_{j=1}^N \beta_{ij} \beta_{hj}}{N \sqrt{\text{Var } \beta_i} \sqrt{\text{Var } \beta_h}} = \frac{\sum_{j=1}^N \alpha_{ij} \beta_{hj}}{N \sqrt{\text{Var } \beta_h}} = \delta_{ih}$$

gelten, verzerrt werden. Aufgabe sei nunmehr, eine lineare Funktion z_1 der x_i so zu finden, daß (6) gilt und der aus den Fehlern β_i der x_i resultierende Anteil an der Varianz von z_1 ein Minimum wird. Wir haben

$$(160) \quad z_1 = \sum_{i=1}^n a'_{1i} x_i$$

und bei Benutzung von (39) sowie (158)

$$(161) \quad V = \text{Var} \beta_1 a'_{11}{}^2 + \text{Var} \beta_2 a'_{12}{}^2 \dots \\ \dots + \text{Var} \beta_n a'_{1n}{}^2 = \text{Var} \beta \sum_{i=1}^n a'_{1i}{}^2,$$

wobei V der zu minimierende Bestandteil der Varianz von z_1 ist ¹⁵³⁾. Da z_1 die Varianz 1 haben soll, ist (161) partiell nach a'_{1i} ($i=1,2,\dots,n$) unter der Nebenbedingung ¹⁵⁴⁾

152) Dabei ist wegen (157)

$$\text{Var } \beta_i = \frac{1}{N} \sum_{j=1}^N \beta_{ij}{}^2.$$

153) GIRSHICK, M.A. (1936), a.a.O., Gleichungen (2.1) und (2.2) auf Seite 523.

154) Vgl. die formale Übereinstimmung mit (151) und (154), Seite 89-90.

$$(162) \quad \text{Var } z_1 = \sum_{i,h=1}^n r_{ih} a'_{1i} a'_{1h} = 1$$

zu differenzieren. Dies führt mit einem LAGRANGESCHEN Multiplikator χ zu

$$(163) \quad \text{Var} \beta \sum_{i=1}^n a'_{1i}^2 - \chi \sum_{i,h=1}^n r_{ih} a'_{1i} a'_{1h}$$

und

$$(164) \quad \text{Var} \beta a'_{1i} - \chi \sum_{h=1}^n r_{ih} a'_{1h} = 0 \quad (i=1,2,\dots,n).$$

Setzen wir

$$(165) \quad \chi = \frac{\text{Var} \beta}{\lambda},$$

so folgt aus (164)

$$(164')$$

$$\sum_{h=1}^n r_{ih} a'_{1h} - \lambda a'_{1i} = 0 \quad (i=1,2,\dots,n).$$

Damit stehen wir, wie bei (150'), erneut vor dem System (53') und müssen wegen ¹⁵⁵⁾

$$(165') \quad v = \frac{\text{Var} \beta}{\lambda}$$

für ein Minimum von (163) den größten Eigenwert λ_1 der Korrelationsmatrix in (164') einsetzen. Auch diese Lösung kann auf die p von Null verschiedenen Eigenwerte ($p \leq n$) von $\mathcal{R}$ erweitert werden, um im allgemeinen Falle

$$(160') \quad z_k = \sum_{i=1}^n a'_{ki} x_i \quad (k=1,2,\dots,p)$$

zu liefern. Nun fehlt aber noch der Beweis, daß die Größen z_k tatsächlich in unserem Sinne Hauptkomponenten sind.

¹⁵⁵⁾ In Analogie zu (151), Seite 89, ergibt sich (165'), wenn man (164) mit a'_{1i} multipliziert und über i summiert.

Hierzu entnehmen wir den Ausführungen von GIRSHICK, daß man mit

$$(166) \quad x_i = \lambda_k \sum_{k=1}^p a_{ki}' z_k \quad (i=1,2,\dots,n)$$

und

$$(167) \quad a_{ik} = \lambda_k a_{ki}'$$

zu den Hauptkomponenten (111'') zurückkehren und damit deren Identität mit (160') zeigen kann ¹⁵⁶⁾.

Der zweite Beitrag von GIRSHICK zur Methodik der Komponentenanalyse besteht in einer Untersuchung von Eigenschaften, die die Hauptkomponenten hinsichtlich ihrer Korrelationen mit den Ausgangsvariablen aufweisen. Sucht man zu den $x_1, x_2, \dots, x_n$ eine Funktion z_1 derart, daß unter der Bedingung (6) die Quadratsumme R_1 der Korrelationskoeffizienten zwischen z_1 und den n Variablen x_i , nämlich - in Analogie zu (8) und (9) -

$$(168) \quad R_1 = \sum_{i=1}^n \left(\frac{1}{N} \sum_{j=1}^N z_{1j} x_{ij} \right)^2 = \sum_{i=1}^n (a_{n1} x_{ih})^2$$

zu einem Maximum wird, so führt die partielle Differentiation unter der angegebenen Nebenbedingung auf einen der Gleichung (164') entsprechenden Ausdruck, mit dessen Hilfe wiederum die zu maximierende Größe als Eigenwert λ_1 von $\mathcal{R}$ identifiziert werden kann ¹⁵⁷⁾. Damit läßt sich beweisen, daß - wie schon HOTELLING es angedeutet hat - die Summe der Quadrate der Korrelationskoeffizienten zwischen der ersten Hauptkomponenten und den Ausgangsvariablen ein Maximum ist. Die Verallgemeinerung dieser Erscheinung auf

$$(169) \quad R_k = \text{Max} ! \quad (k=1,2,\dots,p)$$

156) GIRSHICK, M.A. (1936), a.a.O., Seite 524-525, insbesondere Gleichungen (2.15) und (2.16).

157) GIRSHICK, M.A. (1936), a.a.O., Seite 525-527.

liegt auf der Hand ¹⁵⁸⁾).

An dritter Stelle beschäftigt sich GIRSHICK mit den Konsequenzen, die aus der vorigen Ableitung für die Anwendung der Maximum-Likelihood-Methode auf das Komponentenproblem folgen. Wir kommen auf diese Fragestellung im Zusammenhang mit der mehrdimensionalen Analyse im Abschnitt 303 zurück.

Abschließend sei noch auf die von VAN DEN BOGAARD und KAPTEIN vorgenommene Ableitung zur Komponentenanalyse eingegangen. Diese folgt grundsätzlich der Methode von HOTELLING und GIRSHICK, jedoch werden im Gegensatz zu (5) und (6) die Ausgangsvariablen und Komponenten in der Form

$$(170) \quad \sum_{j=1}^N x_{ij}^2 = \sum_{j=1}^N z_{kj}^2 = 1 \quad (i=1,2,\dots,n; k=1,2,\dots,p),$$

also

$$(171) \quad \bar{x}_i = \bar{z}_k = 0 \quad \text{und}$$

$$(171') \quad \text{Var } \bar{x}_i = \text{Var } \bar{z}_k = \frac{1}{N},$$

normiert. Das hat für die geometrische Interpretation im " n_2 -Raum" die Folge, daß sowohl die Vektoren $\mathcal{P}_i$ als auch die $\mathcal{Z}_k$ Einheitsvektoren sind, und daß die letztgenannten statt einer Orthogonalbasis eine Orthonormalbasis des T_p bilden ¹⁵⁹⁾. Aus diesem Grunde ist es notwendig, für die

158) Vgl. HOTELLING, H. (1933), a.a.O., Seite 422: "An easily verified property of the method is that the first of our principal components has a greater mean square correlation with the tests than any other variable; and that among all variables uncorrelated with the first (p-1) principal components ..., that having the greatest mean square correlation with the tests is the qth principal component." (q=p in unserer Symbolik. Anm.d.Verf.)

159) VAN DEN BOGAARD-KAPTEIN (1958), a.a.O., Seite 1-2. Diese Form der Normierung wird oft als "unitary standard measure" bezeichnet. Vgl. den Gegensatz zu unseren Ausführungen im Abschnitt 12, Seite 6-15. Zu der Normierung nach (171) und (171') siehe auch FRISCH, R. (1934), a.a.O., Seite 11.

z_k nicht die Anteilswerte an den Varianzen, sondern an den Summen der Abweichungsquadrate der x_i , nämlich

$$(170') \quad \sum_{j=1}^N x_{ij}^2 = 1 \quad (i=1,2,\dots,n)$$

zu maximieren ¹⁶⁰⁾. Durch diese Forderung wird der Effekt der Normierung (171) und (171') dahingehend aufgehoben, daß im Endergebnis die Hauptkomponenten die Gestalt (111) annehmen ¹⁶¹⁾. Weiterhin bringen auch VAN DEN BOGAARD und KAPTEIN den ausführlichen Beweis zur Hauptkomponentenlösung nach (168) und (169).

303 Komponentenanalyse und mehrdimensionale Analyse:
 WILKS, T.W. ANDERSON und TINTNER

Von besonderem Interesse in statistischer Hinsicht ist die Komponentenanalyse dann, wenn man die Ausgangsvariablen $x_1, x_2, \dots, x_n$ als Zufallsvariable mit einem n-dimensionalen Verteilungsgesetz deutet. Wenn es auch schwierig sein mag, für das tatsächliche Vorhandensein dieser streng parametrischen Voraussetzung im konkreten Fall einer sozialwissenschaftlichen Fragestellung eine Rechtfertigung zu finden, so bietet andererseits die Unterstellung eines Verteilungsgesetzes die Möglichkeit, exakte Aussagen über einen stichprobentheoretischen Ansatz zur Komponentenanalyse

160) Hierfür verwenden VAN DEN BOGAARD und KAPTEIN (a.a.O., Seite 3) den Ausdruck "Deviance", der von KENDALL vorgeschlagen wird. Vgl. KENDALL-BUCKLAND (1957), a.a.O., Seiten 82 und 358.

161) VAN DEN BOGAARD-KAPTEIN (1958), a.a.O., Gleichung (10.5) auf Seite 14.

zu machen. Das gilt insbesondere bei bestimmten Signifikanzuntersuchungen für Eigenwerte und Hauptkomponenten ¹⁶²⁾.

Im Rahmen der mehrdimensionalen Analyse wird in der Regel nicht von den Korrelationen zwischen den x_i ($i=1,2,\dots,n$), sondern vielmehr von deren Kovarianzen ausgegangen ¹⁶³⁾. Wir fassen zunächst die zufälligen Variablen x_i als Koordinaten eines n -dimensionalen Zufallsvektors (random vector)

$$\tilde{x} = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}$$

auf, für den wegen der Normierung der Variablen ein Mittelwertvektor

$$(172) \quad \mu = E\tilde{x} = 0$$

existiert ¹⁶⁴⁾. Als Kovarianzmatrix wird dann die (n,n) -Matrix

- 162) Aus einer größeren Zahl von Arbeiten auf diesem Gebiet, das wir hier nicht weiter betrachten können, seien folgende genannt:
 GIRSHICK, M.A. (1939) - "On the sampling theory of the roots of determinantal equations", Ann.Math.Stat. 10, Seite 203-224.
 HSU, P.L. (1939) - "On the distribution of roots of certain determinantal equations", Ann.Eugen. 2, Seite 250-258.
 LAWLEY, D.N. (1956) - "Tests of significance for the latent roots of covariance and correlation matrices", Biometrika 43, Seite 128-136.
 PILLAI, K.C.S. (1956) - "On the distribution of the largest or the smallest root of a matrix in multivariate analysis", Biometrika 43, Seite 122-127.
- 163) Wir haben diesen Fall in unserer Ableitung insofern antizipiert, als Korrelationen und Kovarianzen stets identisch sind. Vgl. Fußnote 6 auf Seite 4. Zu weiteren Ausführungen über die Analyse von Kovarianzen anstelle von Korrelationen siehe
 BURT, C. (1937) - "Correlations between persons", Brit.J. Psychol. 28, Seite 59-96, insbesondere Seite 68-70.
- 164) Vgl. dazu
 ANDERSON, T.W. (1958) - An Introduction to Multivariate Statistical Analysis. New York: J. Wiley & Sons (London: Chapman & Hall) Seite 11-19.
 CORSTEN, L.C.A. (1958), a.a.O., Seite 22-24.
 ROPPERT, J. (1959) - Einige Anwendungen des Matrizenkalküls bei der Untersuchung von statistischen Verteilungen. In: Anwendungen der Matrizenrechnung auf wirtschaftliche und statistische Probleme (Einzelschr.d.Deutsch. Stat.Ges. Nr. 9). Würzburg: Physica-Verlag, Seite 216-234.

$$(173) \quad \mathcal{D} = E(\tilde{x} - m)(\tilde{x} - m)' = E(\tilde{x} \tilde{x}')$$

definiert, die in ihrem Aufbau unserer Korrelationsmatrix $\mathcal{R}$ - wenn man die Korrelationen als Kovarianzen auffaßt - entspricht. Die n-dimensionale Normalverteilung der x_i hat eine Wahrscheinlichkeitsdichte ¹⁶⁵⁾

$$(174) \quad \frac{1}{(\sqrt{2\pi})^n \text{Det } \mathcal{D}} \exp \left(-\frac{1}{2} \tilde{x}' \mathcal{D}^{-1} \tilde{x} \right),$$

die im T_p ($p=n$) auf gleichgelagerten Ellipsoiden, die ihren Mittelpunkt im Ursprung haben, der Form

$$(175) \quad \tilde{x}' \mathcal{D}^{-1} \tilde{x} = \frac{1}{\text{Det } \mathcal{D}} \sum_{i,h=1}^n [\text{Kov}(x_i; x_h)] x_i x_h = c \quad (c>0)$$

konstant ist ¹⁶⁶⁾. Die Methode der Komponentenanalyse führt nun wieder auf die Bestimmung der Hauptachsenrichtungen dieser Ellipsoiden mit Hilfe der Eigenwerte der Kovarianzmatrix. Entsprechende Ableitungen finden sich bei WILKS und T.W. ANDERSON ¹⁶⁷⁾, in modifizierter Form unter Berücksichtigung der Ergebnisse von GIRSHICK, jedoch ebenfalls normalverteilte Gesamtheiten voraussetzend auch bei TINTNER und GEARY ¹⁶⁸⁾.

- 165) Wenn für die Kovarianzmatrix eine Kehrmatrix $\mathcal{D}^{-1}$ mit den Elementen $(1/\text{Det } \mathcal{D})\text{Kov}(x_h; x_i) = (1/\text{Det } \mathcal{D})\text{Kov}(x_i; x_h)$ existieren soll, so muß die Verteilung nichtsingulär mit dem Rang n sein. Die vorliegende Ableitung kann auch auf unseren Fall ($p < n$) ausgedehnt werden. Zur singulären Normalverteilung siehe CRAMÉR, H. (1946), a.a.O., Seite 312 und ANDERSON, T.W. (1958), a.a.O., Seite 26-27.
- 166) Vgl. Seite 58, ferner HOTELLING, H. (1933), a.a.O., Seite 427. Die Größe c ist eine beliebige positive Zahl.
- 167) WILKS, S.S. (1950) - Mathematical Statistics. Princeton: Princeton University Press, Seite 252-257.
ANDERSON, T.W. (1958), a.a.O., Seite 278.
- 168) TINTNER, G. (1946) - "Some applications of multivariate analysis to economic data", J.Am.Stat.Ass. 41, Seite 472-500.
- - - Ders. (1950) - "Some formal relations in multivariate analysis", J.Roy.Stat.Soc.B. 12, Seite 95-101.
- - - Ders. (1952), a.a.O., Seite 102-107.
GEARY, R.C. (1948) - "Studies in relations between economic time series", J.Roy.Stat.Soc.B. 10, Seite 140-158.

ANDERSON bringt außerdem - neben einer Erwähnung der Ableitung nach PEARSON¹⁶⁹⁾ - eine Lösung, die von der Definition

"Principal components are linear combinations of random or statistical variables which have special properties in terms of variances"

ausgeht¹⁷⁰⁾ und demzufolge unter Verwendung von (172) sowie (173) zu einer Transformation

$$(176) \quad \tilde{z} = \tilde{e} \eta$$

führt, wobei $\tilde{z}$ der Vektor der Hauptkomponenten mit

$$(177) \quad E(\tilde{z} \tilde{z}') = \mathcal{A}$$

ist¹⁷¹⁾. Schließlich weist ANDERSON darauf hin, daß - wenn man die Ausführungen von (172) bis (177) als Operationen in einer Gesamtheit betrachtet - die aus Stichprobenwerten $x_1, x_2, \dots, x_n$ berechneten Wurzeln der (geschätzten) Kovarianzmatrix $\hat{\sigma}$ und die damit bestimmten Eigenvektoren $\hat{\eta}$ "beste Schätzungen" für die Wurzeln von σ und die Vektoren η im Sinne der Maximum-Likelihood-Methode (maximum likelihood estimation) sind¹⁷²⁾.

169) ANDERSON, T.W. (1958), a.a.O., Seite 278-279. So auch CRAMÉR, H. (1946), a.a.O., Seite 309-310.

170) ANDERSON, T.W. (1958), a.a.O., Seite 272.

171) ANDERSON, T.W. (1958), a.a.O., Theorem (11.2.1) auf Seite 276. $\mathcal{A}$ ist die Kovarianzmatrix der Hauptkomponenten im Sinne von $\mathcal{A}$ auf Seite 56, η entspricht der vollständigen Eigenvektormatrix, deren Spalten η nach (118) normiert sind.

172) ANDERSON, T.W. (1958), a.a.O., Theorem (11.3.1), Seite 279, und Beweis dazu auf Seite 280. Vgl. ferner GIRSHICK, M.A. (1936), a.a.O., Seite 527-528, sowie zur Maximum-Likelihood-Methode ANDERSON, O. (1957), a.a.O., Seiten 126 und 266-267, HOEL, P.G. (1957²⁾) - Introduction to Mathematical Statistics. New York: J.Wiley & Sons (London: Chapman & Hall), Seite 38-40.

31 KOMPONENTENANALYSE UND FAKTORANALYSE

310 Kurzer Abriß der multiplen Faktoranalyse

Im Gegensatz zu der im allgemeinen anwendungsfrei formulierten Komponentenanalyse hat die Faktoranalyse von jeher fast ausschließlich im Zeichen psychologischer Anwendungen gestanden. Sie ist zu Beginn dieses Jahrhunderts aus speziellen Bedürfnissen der Psychologie entstanden¹⁷³⁾ und mit wenigen Ausnahmen bis heute in diesem Anwendungskreis geblieben. In stark vereinfachter Form ist es das Ziel der Faktoranalyse, aus den (Nn) Testergebnissen (test scores), die N Personen in n psychologischen Tests geliefert haben, auf eine Zahl p ($p \leq n$) von zugrundeliegenden Eigenschaften (mental traits) - die Faktoren - zu schließen. Diese Faktoren haben als grundlegende Kategorien der menschlichen

-
- 173) Als Ursprung der Faktoranalyse werden gewöhnlich die Aufsätze von
SPEARMAN, C. (1904a) - "The proof and measurement of association of two things", Am.J.Psychol. 15, Seite 72-101 und
- - - Ders. (1904b) - "'General Intelligence', objectively determined and measured", Am.J.Psychol. 15, Seite 201-293 angegeben. Weitere Quellen zur Herkunft der Faktoranalyse sind:
KRUEGER, F.; SPEARMAN, C. (1906) - "Die Korrelation zwischen verschiedenen geistigen Leistungsfähigkeiten", Zeitschr.f.Psychol. 44, Seite 50-114.
STERN, W. (1917) - Die differentielle Psychologie in ihren methodischen Grundlagen. Leipzig: J.A.Barth, Seite 16-18.
DAMM, H. (1914) - Korrelative Beziehungen zwischen elementaren Vergleichsleistungen (Zeitschr.f.angew. Psychol.u.psychol.Sammelf., Beih. 9). Leipzig: J.A.Barth.
BROWN, W.; THOMSON, G.H. (1925³) - The Essentials of Mental Measurement. Cambridge: At the University Press, Seiten 153-155 und 167-172.
LAZARSFELD, P.F. (1929) - Statistischen Praktikum für Psychologen und Mediziner. Jena: G. Fischer, Seite 168-170.
BURT, C. (1949), a.a.O., Seite 98-104.
VINCENT, D.F. (1953) - "The origin and development of factor analysis", Appl.Stat. 2, Seite 107-117.

Persönlichkeit in dem Sinne zu gelten, daß ihr mehr oder weniger ausgeprägtes Vorhandensein im Individuum dieses zu einer ganz bestimmten Leistung (test performance) im Test x_i ($i=1,2,\dots,n$) befähigt.

Bevor wir zu einer mathematischen Formulierung des Anliegens der Faktoranalyse übergehen, sei zur näheren Erläuterung der oben skizzierten Zielsetzung aus der außerordentlich großen Zahl veröffentlichter Untersuchungen ein neueres Beispiel herausgegriffen. Nach Angabe von BERNYER ¹⁷⁴⁾ wurden zwei Gruppen von Jugendlichen, insgesamt 316 Schüler im Alter von 15 Jahren, 23 Tests unterworfen ($N=316$; $n=23$). Wir deuten diese in der Tabelle 7 kurz an.

Tabelle 7 ¹⁷⁵⁾: Auszug aus einer Testgesamtheit (test battery) nach BERNYER

Test	Bezeichnung des Tests
1	Logical Intelligence
2	Intelligence Test (W)
3	BURT'S Reasoning Test
.	.
.	.
23	Manual Skill

Die Analyse der Testergebnisse führte zur Identifizierung der folgenden Faktoren.

Tabelle 8 ¹⁷⁶⁾: Beschreibung von 5 Faktoren nach BERNYER

Faktor	Identifizierung des Faktors
R	General and Verbal Reasoning
S ₁	Skill in Visualizing a Rigid Structure when moved to different Positions
S ₂	Ability in Visualizing a Configuration in which there is a Movement or Displacement of certain Parts
A	Manual Dexterity
F	Skill in Resolving Problems by concrete Manipulations through Trial and Error

174) BERNYER, G. (1958) - "Second order factors and the organization of cognitive functions", Brit. J. Stat. Psychol. 11, Seite 19-29.

175) BERNYER, G. (1958), a.a.O., Seite 22-23.

176) BERNYER, G. (1958), a.a.O., Seite 21.

Die sachliche Bedeutung dieser Tests und Faktoren interessiert uns hier nicht, ihre Aufzählung mag als vorläufiges Beispiel genügen 177).

Zur einfachen Modellbildung der Faktoranalyse 178) greifen wir weitgehend auf unsere bisherigen Symbole zurück. Es sei $\mathcal{R}$ die Matrix der normierten Testergebnisse von N Personen in n Tests. Sind z_k ($k=1,2,\dots,q>p$) die gesuchten Faktoren 179),

-
- 177) Faktoren können positiv und negativ vorkommen, sich also sachlich in ihr Gegenteil verkehren. Ein drastisches Beispiel hierfür ist ein Faktor "Spiessburger Concernedness versus Bohemian Intellectualism", den wir einer Untersuchung der Persönlichkeitswerte von 373 Studenten entnehmen. Vgl. CATTELL, R.B. (1947) - "Confirmation and clarification of primary personality factors", Psychometrika 12, Seite 216.
- 178) Als Standardwerke zur Faktoranalyse gelten HOLZINGER, K.J.; HARMAN, H.H. (1941) - Factor Analysis. A Synthesis of Factorial Methods. Chicago: The University of Chicago Press. THURSTONE, L.L. (1947) - Multiple-Factor Analysis. Chicago: The University of Chicago Press. CATTELL, R.B. (1952) - Factor Analysis: An Introduction and Manual for the Psychologist and Social Scientist. New York: Harper & Brothers. FRUCHTER, B. (1954) - Introduction to Factor Analysis. New York: D. Van Nostrand Co. Diese Aufstellung ist durch drei weitere Titel zu ergänzen, die trotz der Bemühungen des Verfassers in Deutschland nicht zu erhalten waren: KELLEY, T.L. (1935) - Essential Traits of Mental Life (Harvard Studies in Education, Vol. 26). Cambridge: Harvard University Press. BURT, C. (1940) - The Factors of Mind. London: The University of London Press. THOMSON, G.H. (1951^{5a}) - The Factorial Analysis of Human Ability. London: The University of London Press.
- 179) Wir betrachten hier nur die von THURSTONE eingeführte multiple Faktoranalyse oder Mehrfaktorenanalyse (multiple factor analysis). Ursprünglich konzentrierte sich die Faktoranalyse auf einen Intelligenzfaktor (SPEARMAN "G") bzw. auf zwei Faktoren (HOLZINGERS Bifaktortheorie). Vgl. SPEARMAN, C. (1904b), a.a.O., Seite 202. HOLZINGER, K.J.; SWINEFORD, F. (1937) - "The bi-factor method", Psychometrika 2, Seite 41-54. HENRYSSON, S. (1957) - Applicability of Factor Analysis in the Behavioral Sciences. A Methodological Study (Stockholm Studies in Educational Psychology I). Stockholm: Almqvist & Wiksell, Seite 32-40.

so gelten unter den gleichen Voraussetzungen wieder

$$(28''') \quad \mathcal{H} = \alpha \mathcal{Z},$$

$$(30') \quad \mathcal{R} = \frac{1}{N} \mathcal{H} \mathcal{H}'$$

und

$$(36'') \quad \mathcal{R} = \alpha \alpha'.$$

Die Matrix α ist jetzt als Faktormatrix zu bezeichnen, ihre Elemente heißen Faktorladungen (factor loadings, test coefficients, reference abilities); da sie die "Gewichte" darstellen, mit denen die Faktoren nach einer Beziehung vom Typ (11) an jedem Test beteiligt sind ¹⁸⁰⁾, kommt ihnen in der Faktoranalyse eine weitaus größere Bedeutung als den Faktoren z_k - die auf ein Individuum j bezogen sind - zu ($j=1,2,\dots,N$). Im Zusammenhang mit dem Zahlenbeispiel zur Komponentenanalyse hatten wir bereits angedeutet ¹⁸¹⁾, daß

- 180) Diese Fragestellung ist umkehrbar. Hier werden Tests miteinander korreliert, um die der Testgesamtheit zugrundeliegende Faktorstruktur zu ermitteln. Man kann auch die Personen korrelieren (also nicht die Zeilen von $\mathcal{H}$, sondern die Spalten) und damit das unterschiedliche Verhalten innerhalb der jeweiligen Personengesamtheit untersuchen. Den ersten Fall, mit dem wir es hier ausschließlich zu tun haben, nennt man "R-Technik", den zweiten "Q-Technik". Siehe STEPHENSON, W. (1935) - "Technique of factor analysis", Nature 136, Seite 297.
 - - - - Ders. (1936) - "Introduction to inverted factor analysis, with some applications to studies in orexis", J. Educ. Psychol. 27, Seite 353-367.
 BURT, C. (1937), a.a.O., Seite 59-96.
 CATTELL, R.B. (1952), a.a.O., Seite 88-107.
 BASS, B.M. (1957) - "Iterative inverse factor analysis - a rapid method for clustering persons", Psychometrika 22, Seite 105-107.
 Nur angedeutet sei, daß es noch weitere Untersuchungsrichtungen gibt, so zum Beispiel die Anwendung der Faktoranalyse für das Verhalten einer Person im Zeitablauf ("O-Technik"). Vgl.
 BURT, C.; WATSON, H. (1951) - "Factor analysis of assessments for a single person", Brit. J. Psychol. Stat. Sect. 4, Seite 179-191.

- 181) Vgl. Abschnitt 213, Seite 80-83.

die Struktur der Matrix α - insbesondere die der Quadrate ihrer Elemente, wie aus den nachfolgenden Gleichungen (179) bis (180) hervorgehen wird - bei der Faktoranalyse als Ausdruck der Beziehung zwischen Tests und Faktoren in den Vordergrund tritt, während in der Komponentenanalyse das größere Gewicht auf der Matrix β liegt. Man unterscheidet in der Faktoranalyse mehrere Arten von Faktoren, nämlich die gemeinsamen Faktoren (common factors), die spezifischen Faktoren (specific factors) und die Fehlerfaktoren (error factors). Gemeinsame Faktoren ¹⁸²⁾ treten gleichzeitig bei mehreren Tests auf, während jedem Test mit Sicherheit ein Fehler- oder Restglied, darüber hinaus in den meisten Fällen ein spezifischer Faktor eigentümlich ist. Anstelle von (11) haben wir also von einem Modell

$$(178) \quad x_i = \sum_{k=1}^q a_{ik} z_k = \sum_{k=1}^p A_{ik} f_k + \sum_{k=p+1}^t B_{ik} s_k + C_{ii} e_i$$

(i=1,2,...,n)

auszugehen, das die f_k (k=1,2,...,p) als gemeinsame Faktoren, die spezifischen Faktoren s_k (k=p+1,p+2,...,t) sowie die Restgröße e_i - alle mit den entsprechenden Faktorladungen A_{ik} , B_{ik} und C_{ii} versehen - enthält. Unter diesen Umständen ist α eine $[n,q=n(t+1)]$ -reihige Matrix ¹⁸³⁾, bei der für einige der Elemente ($a_{ik}=0$) gilt.

$$\alpha = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1p} & B_{11} & \dots & B_{1t} & 0 & \dots & 0 & 0 & \dots & 0 & C_{11} & 0 & \dots & 0 \\ A_{21} & A_{22} & \dots & A_{2p} & 0 & \dots & 0 & B_{21} & \dots & B_{2t} & 0 & \dots & 0 & 0 & C_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \dots & A_{np} & 0 & \dots & 0 & 0 & \dots & 0 & B_{n1} & \dots & B_{nt} & 0 & 0 & \dots & C_{nn} \end{bmatrix}.$$

Die Korrelationskoeffizienten zwischen den Variablen x_i sind nach (36'') wie folgt zusammengesetzt:

182) Hier trennt man weiter in General- (general factors) und Gruppenfaktoren (group factors), je nachdem, ob ein gemeinsamer Faktor in allen Tests einer Ausgangsgesamtheit oder nur in einer Gruppe von ihnen vorkommt.

183) Aus formalen Gründen nehmen wir an, daß die Zahl der spezifischen Faktoren bei allen Tests dieselbe ist. Tatsächlich ist diese Voraussetzung nicht erforderlich.

$$(179) \quad r_{ih} = \begin{cases} \text{Var } x_i = \sum_{k=1}^p A_{ik}^2 + \sum_{k=p+1}^t B_{ik}^2 + C_{ii}^2 & \text{für } i=h \\ \text{Kov } (x_i; x_h) = \sum_{k=1}^p A_{ik} A_{hk} & \text{für } i \neq h. \end{cases}$$

(i, h=1, 2, ..., n)

Wir sehen, daß die Faktorladungen der spezifischen Faktoren und des Fehlerfaktors nur in den Varianzen der x_i , nicht aber in deren Kovarianzen vorkommen; sie betreffen, mit anderen Worten, nur die Hauptdiagonale der Korrelationsmatrix, nicht die Koeffizienten r_{ih} ($i \neq h$).

An der Gleichung

$$(179') \quad \text{Var } x_i = \sum_{k=1}^p A_{ik}^2 + \sum_{k=p+1}^t B_{ik}^2 + C_{ii}^2 = 1 \quad (i=1, 2, \dots, n)$$

lassen sich einige wichtige Begriffe der Faktoranalyse erläutern. Der Ausdruck ¹⁸⁴⁾

¹⁸⁴⁾ Im folgenden werden die in der englischsprachigen Literatur fest eingepprägten Begriffe benutzt, weil die Schreibweise im Deutschen uneinheitlich ist. Für (180), Seite 107, findet man z.B. "A-Streuung" (GRÜNEBERG), "Gemeinsamkeit" (HOFSTÄTTER), "Gemeinsamkeitsgrad" oder "Varianz der gemeinsamen Faktoren" (KENDALL-BUCKLAND) und "Kommunalität" (LIENERT). Vgl. GRÜNEBERG, H.-J. (1952) - "Die multiple Faktoranalyse", Mittbl. Math. Stat. 4, Seite 14. HOFSTÄTTER, P.R. (1953) - Einführung in die quantitativen Methoden der Psychologie. München: J.A. Barth, Seite 140. KENDALL-BUCKLAND (1957), a.a.O., Seiten 367 und 370. LIENERT, G.A. (1959) - "Prinzip und Methode der multiplen Faktorenanalyse, demonstriert an einem Beispiel", Biom. Zeitsch. 1, Seite 95. Über (180) und (180') hinaus werden auch die Begriffe $\sum_{k=1}^t A_{ik}^2 + B_{ik}^2$ = "reliability" (Verläßlichkeit), $\sum_{k=p+1}^t B_{ik}^2$ = "specificity" und C_{ii}^2 = "unreliability" verwendet.

$$(180) \quad h_i^2 = \sum_{k=1}^p A_{ik}^2 \quad (i=1,2,\dots,n),$$

der den Anteil der gemeinsamen Faktoren f_k ($k=1,2,\dots,p$) an der Varianz der Tests x_i wiedergibt, wird als "communality" bezeichnet, die für praktische Zwecke untrennbare Restgröße

$$(180') \quad u_i^2 = 1 - h_i^2 = \sum_{k=p+1}^t B_{ik}^2 + C_{ii}^2 \quad (i=1,2,\dots,n)$$

heißt "uniqueness" von x_i .

Aus (178) folgt, daß in der Modellvorstellung der Faktoranalyse die Zahl der Faktoren in jedem Falle größer als die der Tests ist. Das Gleichungssystem ist unbestimmt, die Zahl der Gleichungen ist kleiner als die der unbekannten Größen. Man muß sich auf einen Teil der Faktoren beschränken und arbeitet aus diesem Grunde in der Faktoranalyse stets nur mit den gemeinsamen Faktoren $f_1, f_2, \dots, f_p$ ($p \leq n$). Dies wird erreicht, indem aus der "vollständigen" Korrelationsmatrix (complete correlation matrix) $\mathcal{R}$ der "Einfluß" nicht-gemeinsamer Faktoren durch

$$(181) \quad r_{ii} - u_i^2 = h_i^2 \quad (i=1,2,\dots,n)$$

eliminiert wird. Das bedeutet, daß wegen (179) in der Hauptdiagonalen von $\mathcal{R}$ die Korrelationskoeffizienten r_{ii} durch die Summen der Quadrate der Faktorladungen der gemeinsamen Faktoren ersetzt werden. Wir haben dann eine "reduzierte" Korrelationsmatrix (reduced correlation matrix) $\mathcal{R}_{(r)}$, die mit der ebenfalls verkürzten Faktormatrix

$$\mathcal{F} = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1p} \\ A_{21} & A_{22} & \dots & A_{2p} \\ \dots & \dots & \dots & \dots \\ A_{n1} & A_{n2} & \dots & A_{np} \end{bmatrix}$$

das zwar nicht eindeutige, aber lösbares Faktorsystem

$$(36''') \quad \mathcal{R}_{(r)} = ff'$$

bildet ¹⁸⁵⁾.

Wir kehren nun nochmals zu der am Anfang dieses Abschnitts als Beispiel zitierten Analyse von BERNYER zurück und bringen zur Verdeutlichung der bisherigen Ausführungen im Auszug die ersten beiden Zeilen der Faktormatrix f :

Tabelle 9 ¹⁸⁶⁾: Faktorladungen von 5 gemeinsamen Faktoren nach BERNYER

Test	Faktorladungen A_{ik} ($i=1,2;k=1,2,3,4,5$) der gemeinsamen Faktoren					h_i^2 ($i=1,2$)
	R	S_1	S_2	A	F	
1	0,58	-0,03	0,18	0,05	-0,03	0,3731
2	0,57	0,03	0,07	-0,05	-0,05	0,3357
.	.	.	.	.	.	.

Betrachten wir die erste Zeile der Faktormatrix, so haben wir laut (178) für den Test 1 (Logical Intelligence) eine Beziehung

$$(178') \quad x_1 = 0,58f_{(R)} - 0,03f_{(S_1)} + 0,18f_{(S_2)} + 0,05f_{(A)} - 0,03f_{(F)},$$

die mit

$$(180'') \quad h_1^2 = \sum_{k=1}^5 A_{1k}^2 = 0,3731$$

37,31 % der Varianz von x_1 "erklärt", während im Einzelfall zum Beispiel die Leistung im Test 1 durch $A_{11}^2 = 0,3364$ mit

185) Zur Mehrdeutigkeit der Gleichung (36''') vgl. Seite 20-21. Diese Gleichung nimmt in der Faktoranalyse eine zentrale Stellung ein, als "fundamental equation of factor analysis" ist sie in allen Abhandlungen der Methode anzutreffen. So zum Beispiel THURSTONE, L.L. (1947), a.a.O., Seite 79.

186) BERNYER, G. (1958), a.a.O., Seite 22. Zur Ergänzung sei gesagt, daß die Faktorladungen der Tabelle 9 nach der Schwerpunktmethode mit anschließender Bewegung auf die Einfachstruktur bestimmt wurden. Das Prinzip der Einfachstruktur wird weiter unten auf den Seiten 112 und 117-118 erläutert. Zur Erklärung der Tests und Faktoren siehe Tabellen 7 und 8 auf Seite 102.

33,64 % durch den als "General and Verbal Reasoning" gekennzeichneten Faktor R beeinflusst wird. Wir begnügen uns mit dieser kurzen Deutung der Faktorladungen und verweisen erneut auf die Analogie zu den Tabellen 5 und 6¹⁸⁷⁾.

Mit den Gleichungen (181) und (36''') haben wir den Punkt berührt, der die Faktoranalyse von der Komponentenanalyse trennt und der zugleich das schwächste, wenn nicht überhaupt bedenkliche unter den methodischen Elementen der Faktoranalyse ist. Dazu muß man sich vor Augen halten, daß die Größen h_i^2 in jeder Analyse dann, wenn sie zur Reduktion der Korrelationsmatrix benötigt werden, grundsätzlich unbekannt sind. Nach (180) setzen sie die Kenntnis der Zahl und der unterschiedlichen Einwirkung der gemeinsamen Faktoren voraus, während doch schließlich die Bestimmung dieser Daten das Endziel der Analyse bildet. Eine Auflösung dieses Widerspruchs ist auf zwei Wegen denkbar.

Man kann einmal von einer Hypothese darüber ausgehen, wie groß in der zu untersuchenden Gesamtheit die Zahl der gemeinsamen Faktoren sein wird und dann mit Hilfe bestimmter Schätzverfahren die Werte h_i^2 mit dieser Hypothese abstimmen. Eine solche Fragestellung führt über den Begriff der "Faktorstruktur" auf das tiefer liegende Problem des "Wesens" der Faktoren überhaupt - etwa im Gegensatz zu dem der Komponenten z_k -, das wir erst im Abschnitt 311 aufgreifen werden. Der andere Weg geht von der Forderung aus, die Zahl p der gemeinsamen Faktoren möglichst klein zu halten¹⁸⁸⁾. Da p nach der Definition unkorrelierter Faktoren als Rang von f und damit wegen (36''') auch als Rang von $R_{(r)}$ zu

187) Vgl. Seite 81.

188) Diese Forderung ist von THURSTONE zu einem der grundlegenden Prinzipien der Faktoranalyse formalisiert worden, das man als "Sparsamkeit in der Beschreibung einer Gesamtheit" (economy or parsimony in description) bezeichnen kann. Vgl. auch FERGUSON, G.A. (1954) - "The concept of parsimony in factor analysis", Psychometrika 19, Seite 281-290.

betrachten ist, die vollständige Korrelationsmatrix $\mathcal{R}$ aber einen durch die Tests vorgegebenen Rang besitzt, wäre die Zahl der Faktoren durch Einsetzen von solchen Werten h_i^2 in die Hauptdiagonale von $\mathcal{R}_{(r)}$ möglichst klein zu halten, die den Rang p von $\mathcal{R}_{(r)}$ zu einem Minimum machen. Dieses Vorgehen ist im Gegensatz zu dem der Abschätzung auf Grund einer Hypothese objektiv; es besteht darin, zu den gegebenen Elementen r_{ih} ($i \neq h$) einer symmetrischen (n,n) -Matrix für die Hauptdiagonale n Werte h_i^2 so zu finden, daß

$$(182) \quad 0 \leq h_i^2 \leq 1 \quad (i=1,2,\dots,n)$$

und für den Rang

$$(183) \quad p = \text{Min} !$$

gelten. Die Lösung der Aufgabe (183) unter der Bedingung (182) liegt bisher nicht in befriedigender Form vor. Man kann obere und untere Grenzen für den kleinstmöglichen Rang p angeben, ohne allerdings innerhalb dieser einer eindeutigen Reihe $h_1^2, h_2^2, \dots, h_n^2$ und vor allem der Einhaltung von (182) gewiß zu sein. Nach LEDERMANN¹⁸⁹⁾ ist es unter den angedeuteten Vorbehalten möglich, den Rang der Korrelationsmatrix auf die der Größe

$$\frac{1}{2} \left[(2n + 1) - \sqrt{8n + 1} \right]$$

nächstliegende ganze Zahl zu reduzieren, während GUTTMAN gezeigt hat¹⁹⁰⁾, daß die untere Grenze von p durch die Zahl der positiven Eigenwerte einer Korrelationsmatrix gegeben ist, die in der Hauptdiagonalen statt der Elemente r_{ii} die Quadrate der multiplen Korrelationskoeffizienten zwischen den Tests enthält. Zur ausführlichen Diskussion des minimalen

189) LEDERMANN, W. (1937) - "On the rank of the reduced correlational matrix in multiple factor analysis", Psychometrika 2, Seite 85-93. Vgl. auch THURSTONE, L. L. (1947), a.a.O., Seite 292-294.

190) GUTTMAN, L. (1954a) - "Some necessary conditions for common-factor analysis", Psychometrika 19, Seite 149-161.

Ranges von $R_{(r)}$ müssen wir auf eine Reihe von speziellen Arbeiten verweisen ¹⁹¹⁾.

Nach Vorgabe der Varianzen der gemeinsamen Faktoren und Einsetzen derselben in die Hauptdiagonale der Korrelationsmatrix kann man dazu übergehen, die Faktorladungen A_{ik} zu bestimmen. Hierfür gibt es eine Reihe von Lösungsverfahren, in der Praxis wird jedoch fast ausschließlich die von THURSTONE entwickelte und von uns bereits erwähnte ¹⁹²⁾ Schwerpunkts- oder Zentroidmethode verwendet. Wir werden dieses Verfahren im Anschluß an die nun folgenden Ausführungen zur Anwendung von geometrischen Analogien in der Faktoranalyse kurz besprechen.

Die geometrische Anschauung der Faktoranalyse findet grundsätzlich im " n_2 -Raum" des Abschnitts 301 statt ¹⁹³⁾. Wir können daher unter bestimmten Einschränkungen erneut die Vorstellungen verwenden, die wir der Entwicklung der Komponentenanalyse zugrundegelegt hatten. Die Tests bilden demnach ein Gebilde von Radiusvektoren, für deren Winkel

-
- 191) ECKART, C.; YOUNG, G. (1936) - "The approximation of one matrix by another of lower rank", Psychometrika 1, Seite 211-218.
LEDERMANN, W. (1940) - "A problem concerning matrices with variable diagonal elements", Proc. Roy. Soc. Edinb. 60, Seite 1-17.
ALBERT, A. A. (1944a) - "The matrices of factor analysis", Proc. Nat. Acad. Sci. USA. 30, Seite 90-95.
- - - Ders. (1944b) - "The minimum rank of a correlation matrix", Proc. Nat. Acad. Sci. USA. 30, Seite 144-146.
ROSNER, B. (1948) - "An algebraic solution for the communalities", Psychometrika 13, Seite 181-184.
GUTTMAN, L. (1958) - "To what extent can communalities reduce rank?", Psychometrika 23, Seite 297-308.
WRIGLEY, C. R. (1959) - "The effect upon the communalities of changing the estimate of the number of factors", Brit. J. Stat. Psychol. 12, Seite 35-51.

192) Vgl. Seite 62.

193) Vgl. Seite 90-91.

die Beziehung (19) gilt. Da nur die gemeinsamen Faktoren interessieren (common factor space), hat das Vektorgebilde - auch Konfiguration (configuration of test vectors) genannt - die Dimension p ; die Länge der Vektoren ist dadurch, im Gegensatz zu der der Vektoren $\mathcal{P}_i$, durch die Größen h_i ($i=1,2,\dots,n$) gegeben. Die p Faktoren sind als ein System von orthogonalen Einheitsvektoren (reference vectors) mit dem gleichen Ursprungspunkt wie die Testvektoren zu deuten, die Faktorladungen schließlich als die Projektionen der Testvektoren auf die Faktoren. Während die Komponentenanalyse die Mehrdeutigkeit einer Beziehung vom Typ (36''') durch die Einführung von Hauptkomponenten beseitigt, wird in der Faktoranalyse die Eindeutigkeit der Lösung im allgemeinen mit Kriterien hergestellt, die außerhalb eigentlich methodischer Erwägungen liegen. Zunächst bezeichnet man jede einmal gegebene relative Lage zwischen der Konfiguration der Testvektoren und dem System der Faktoren als eine Faktorstruktur (factor structure). Eine solche kann durch Bewegung der Achsen so verändert werden, daß die Faktorladungen eine aus sachlichen Gründen her wünschenswerte Ausprägung erhalten. Als speziell zu bevorzugende Struktur wird diejenige angesehen, bei der möglichst viele Faktorladungen A_{ik} den Wert Null annehmen und möglichst wenige ein negatives Vorzeichen besitzen ¹⁹⁴⁾. Eine Struktur dieser Art heißt "Einfachstruktur" (simple structure), ihre Bedeutung liegt, wie wir noch sehen werden, in einer besonders günstigen Deutungsmöglichkeit der Faktoren. Die Einfachstruktur kann durch graphische Methoden approximiert, aber auch unter Verwendung bestimmter Minimumsbedingungen direkt auf analytischem Wege berechnet werden. Es ist weiterhin möglich, zum Erreichen der Einfachstruktur das Prinzip

194) Die Kriterien werden hier etwas vereinfacht wiedergegeben. Für weitere Einzelheiten vgl. THURSTONE, L.L. (1947), a.a.O., Seite 335. LIENERT, G.A. (1959), a.a.O., Seite 106-108.

orthogonaler Faktoren aufzugeben und korrelierte Faktoren (oblique factors) zuzulassen ¹⁹⁵⁾. Die Schwerpunktmethode geht wie die Methode der Hauptkomponenten von der Absicht aus, die Faktorladungen der aufeinanderfolgenden Faktoren f_k ($k=1,2,3,\dots,p$) möglichst groß zu halten. Sie werden aber nicht im strengen Sinne maximiert; vielmehr folgt man der Überlegung, daß die Projektionen einer Konfiguration von Testvektoren auf eine Achse dann besonders groß sind, wenn diese durch die Mitte der Konfiguration hindurchläuft. Dementsprechend werden die Faktorladungen A_{i1} ($i=1,2,3,\dots,n$) so eingerichtet, daß der Faktorvektor $\mathbf{z}_1$ durch den Schwerpunkt der Konfiguration gelegt wird. Leider läßt sich dieses Verfahren für den Vektor $\mathbf{z}_2$ schon nicht mehr anwenden,

-
- 195) Zu weiteren Ausführungen über die Einfachstruktur und die verschiedenen Berechnungsverfahren siehe THURSTONE, L.L. (1940) - "Experimental study of simple structure", Psychometrika 5, Seite 153-168. HORST, P. (1941) - "A non-graphical method for transforming an arbitrary factor matrix into a simple structure", Psychometrika 6, Seite 79-99. TUCKER, L.R. (1944b) - "A semi-analytical method of factorial rotation to simple structure", Psychometrika 9, Seite 43-68. ZIMMERMANN, W.S. (1946) - "A simple graphical method for orthogonal rotation of axes", Psychometrika 11, Seite 51-55. HARRIS, C.W.; KNOELL, D.M. (1948) - "The oblique solution in factor analysis", J. Educ. Psychol. 39, Seite 385-403. REYBURN, H.A.; RAATH, M.J. (1949) - "Simple structure: A critical examination", Brit. J. Psychol. Stat. Sect. 2, Seite 125-135. SUTHERLAND, J. (1951) - "Factor rotation by the method of extended vectors", Brit. J. Psychol. Stat. Sect. 4, Seite 21-30. GIBSON, W.A. (1952) - "Orthogonal and oblique simple structures", Psychometrika 17, Seite 317-323. CARROLL, B. (1953) - "An analytical solution for approximating simple structure in factor analysis", Psychometrika 18, Seite 23-38. THURSTONE, L.L. (1954) - "An analytical method for simple structure", Psychometrika 19, Seite 173-182. NEUHAUS, J.O.; WRIGLEY, C. (1954) - "The quartimax method. An analytic approach to orthogonal simple structure", Brit. J. Stat. Psychol. 7, Seite 81-91. TUCKER, L.R. (1955) - "The objective definition of simple structure in linear factor analysis", Psychometrika 20, Seite 209-225.

weil dann der Schwerpunkt der Residualvektoren 196) im Nullpunkt liegt, so daß ohne Zusatzmaßnahmen keine weiteren Faktorladungen berechnet werden können. Man hilft sich durch Vorzeichenwechsel in der reduzierten Korrelationsmatrix $R_{.1(r)}^*$, wodurch der Schwerpunkt verschoben wird. Der Wechsel von Vorzeichen bestimmter Elemente der Matrix ist eine Angelegenheit subjektiver Entscheidung, so daß die Schwerpunktmethode - ganz abgesehen von den in der Matrix $R_{(r)}$ befindlichen "communalities" - nicht die Exaktheit der Komponentenanalyse aufweist. Immerhin ist der erforderliche Rechenaufwand geringer, auch lassen sich die Schwerpunktfaktoren (centroid factors) als durch Iteration verbessereungsfähige Approximationen der Hauptkomponenten auffassen 197).

Wir haben damit in aller Kürze einige Grundprobleme der Faktoranalyse behandelt. Während wir auf die spezielle Frage des Zusammenhangs zwischen Komponenten- und Faktoranalyse noch eingehen werden, müssen wir auf die Erörterung tiefer gehender Probleme der Faktoranalyse, insbesondere solcher ihrer neueren Entwicklung 198), verzichten. Die Literatur

-
- 196) Diese hatten wir im Abschnitt 160, Seite 42-45, mit $\mathcal{C}_{i.1}^*$, allgemein mit $\mathcal{C}_{i.k-1}^*$ bezeichnet. Vgl. auch (96) auf Seite 49. Näheres dazu siehe bei KENDALL, M.G. (1957), a.a.O., Seite 27-34.
- 197) BARGMANN, R. (1956), a.a.O., Seite 1-14.
- 198) Darunter sind in erster Linie die von LAZARSELD und GUTTMAN ausgehenden Bestrebungen zu verstehen. Siehe GREEN, B.F. (1952) - "Latent structure analysis and its relation to factor analysis", J. Am. Stat. Ass. 47, Seite 71-76.
- LAZARSELD, P.F. (1954) - A Conceptual Introduction to Latent Structure Analysis. In: LAZARSELD, P.F. (Hrsg.) - Mathematical Thinking in the Social Sciences. Glencoe: The Free Press, Seite 349-387.
- GUTTMAN, L. (1954b) - A New Approach to Factor Analysis: The Radex. In: LAZARSELD, P.F. (Hrsg.), a.a.O., Seite 258-348.
- - - Ders. (1955) - "A generalized simplex for factor analysis", Psychometrika 20, Seite 173-192.
- ANDERSON, T.W. (1959) - Some Scaling Models and Estimation Procedures in the Latent Class Model. In: GREENANDER, U. (Hrsg.) - Probability and Statistics. The HARALD CRAMER Volume. Stockholm: Almqvist & Wiksell (New York: Wiley & Sons), Seite 9-38.

zur Faktoranalyse ¹⁹⁹⁾ hat einen Umfang angenommen, der eine Darstellung aller Einzelfragen als über den letztlich an der Komponentenanalyse orientierten Rahmen unserer Betrachtung hinausgehend verbietet ²⁰⁰⁾.

311 Über die Unterscheidung zwischen Komponentenanalyse und Faktoranalyse

Aus dem Anlaß einer ersten Übersicht über die Entwicklung der Komponentenanalyse ²⁰¹⁾ hatten wir bereits festgestellt, daß HOTELLING bei der Darstellung seiner Methode von ihrer Anwendung im Rahmen der Faktoranalyse ausgegangen war. Auch in der Folgezeit ist die Methode der Komponentenanalyse stets als ein spezielles Verfahren der Faktoranalyse im Sinne der Aufgabenstellung des Abschnitts 310 betrachtet und in der Fachliteratur entsprechend

199) So verzeichnet FRUCHTER (1954, a.a.O., Seite 221-266) insgesamt 688 Titel, die sich mit einigen Ausnahmen lediglich auf die Zeit von 1940 bis 1955 beziehen.

200) Neben den bereits in der Fußnote 184, Seite 106, angegebenen Veröffentlichungen von GRÜNEBERG, HOFSTÄTTER und LIENERT sind an Einführungen in deutscher Sprache zu nennen:
HOFSTÄTTER, P.R. (1938) - "Über Faktoren-Analyse", Arch. f.d.ges.Psychol. 100, Seite 223-279.
GOTSCHALDT, K. (1940) - Die Statistik in der Psychologie. In: BURGDORFER, F. (Hrsg.) - Die Statistik in Deutschland nach ihrem heutigen Stand. Ehrengabe für FRIEDRICH ZAHN. Berlin: Verlag f. Sozialpol., Wirtsch. u. Stat. P. Schmidt, Seite 392-393.
TOMAN, W. (1949) - "Die Faktorenanalyse in der Psychologie", Stat. Vierteljahresschr. 2, Seite 57-68.
Eine elementare Einführung in englischer Sprache bieten PETERS, C.C.; VAN VOORHIS, W.R. (1940) - Statistical Procedures and their Mathematical Bases. New York (London): McGraw Hill Book Co., Seite 248-278.
QUENOUILLE, M.H. (1952) - Associated Measurements. London: Butterworths Scientific Publ., Seite 199-203.
ADCOCK, C.J. (1954) - Factorial Analysis for Non-Mathematicians. Carlton: Melbourne University Press.
JOWETT, G.H. (1958) - "Factor analysis", Appl. Stat. 7, Seite 114-125.

201) Vgl. Seite 1.

behandelt worden ²⁰²⁾. Die von uns im Abschnitt 1 dargestellte Theorie kann auch auf einen Ansatz der Form (178) angewendet werden, nämlich dann, wenn nur eine Analyse des von den gemeinsamen Faktoren aufgespannten Raumes beabsichtigt ist - die Überführung in das Eigenwertproblem der Korrelationsmatrix $\mathcal{R}$ kann man ohne Schwierigkeiten auf die reduzierte Korrelationsmatrix $\mathcal{R}_{(r)}$ ausdehnen; die gemeinsamen Faktoren f_k sind dann in unserem Sinne Hauptkomponenten, deren anschließender Bewegung auf eine zu bevorzugende Faktorstruktur nichts im Wege steht ²⁰³⁾. Wir haben dieses Vorgehen durch die Gleichung (28''') bereits impliziert.

-
- 202) In der Tat verwendet HOTELLING den seither eingebürgerten Begriff der "Komponenten" beziehungsweise "Hauptkomponenten" nur deshalb, weil nach seiner Ansicht der Ausdruck "Faktor" zu leicht mit dem mathematischen Begriff des Faktors verwechselt werden kann. Vgl. HOTELLING, H. (1933), a.a.O., Seite 417.
- 203) Die Kontroversen über das Verfahren von HOTELLING betreffen in erster Linie die Frage, ob von der Matrix $\mathcal{R}$ oder von $\mathcal{R}_{(r)}$ ausgegangen werden soll. Dies hat sachliche, nicht methodische Gründe - das Verfahren selbst kann in beiden Fällen verwendet werden. Vgl. etwa KELLOGG, C.E. (1936a) - "The problem of principal components: Derivation of HOTELLING'S method from THURSTONE'S", J.Educ.Psychol. 27, Seite 512-520.
 - - - Ders. (1936b) - "The problem of principal components: II - The argument for communalities", J.Educ. Psychol. 27, Seite 581-590.
 McCLOY, C.H.; METHENY, E.; KNOTT, V. (1938) - "A comparison of the THURSTONE method of multiple factors with the HOTELLING method of principal components", Psychometrika 3, Seite 61-67.
 CURETON, E. (1939) - "The principal compulsions of factor-analysts", Harv.Educ.Rev. 9, Seite 287-295.
 HOLZINGER, K.J. (1946) - "A comparison of the principal-axes and centroid factors", J.Educ.Psychol. 37, Seite 449-472.
 Es sei angemerkt, daß die Komponentenanalyse als Verfahren der Faktoranalyse gewöhnlich den Namen "Hauptachsenmethode" trägt. Während ihre Entwicklung in aller Regel HOTELLING und seit einer entsprechenden Bemerkung von BURT auch PEARSON zugeschrieben wird, hat gleichwohl THURSTONE bereits 1932 die Methode der Hauptachsen empfohlen. Vgl.
 THURSTONE, L.L. (1944) - "Graphical method of factoring the correlation matrix", Proc.Nat.Acad.Sci. USA. 30, Seite 129.
 BURT, C. - Diskussionsbeitrag zu KENDALL-BABINGTON SMITH (1950), a.a.O., Seite 87-88.

Die Unterscheidung zwischen Komponentenanalyse und Faktoranalyse, die wir jetzt untersuchen wollen, ist weniger durch methodische Abweichungen als vielmehr durch gewisse Annahmen über die Natur der berechneten Komponenten oder Faktoren gekennzeichnet. Während es in der Komponentenanalyse darauf ankommt, einen gegebenen Komplex von Ausgangsvariablen unter dem Gesichtspunkt maximaler Erfassung der Gesamtvariation durch eine möglichst geringe Zahl von Hauptkomponenten darzustellen - wobei eine sachliche Bedeutung derselben möglich, nicht aber als Kriterium ihrer Gültigkeit aufzufassen ist -, geht die Faktoranalyse von einer aus sachlichen Gründen her postulierten Faktorstruktur aus, die durch die Analyse verifiziert werden soll ²⁰⁴⁾. Hauptkomponenten sind, mit anderen Worten, optimale mathematische Größen (mathematical artefacts), Faktoren dagegen numerischer Ausdruck von effektiven Einflußgrößen (real entities), die die Gestalt einer Ausgangsgesamtheit gewissermaßen "physisch" geformt haben - wie es etwa in dem Zusammenspiel der Tests und Faktoren der von uns einleitend betrachteten Untersuchung von BERNYER ²⁰⁵⁾ der Fall ist. In diesem Sinne ist auch das Prinzip der Einfachstruktur und das der Bewegung der Faktorachsen überhaupt zu verstehen. Es kommt in der Faktoranalyse nicht nur darauf an, die bestmögliche Verringerung der Dimension der Ausgangsgesamtheit zu erreichen, sondern darüber hinaus eine befriedigende Erklärung der Faktoren (psychological meaningful results) zu finden. Da negative Faktorladungen diese Interpretation durch die dann erforderliche Einführung von bipolaren Faktoren erschweren, andererseits ein Faktor desto leichter zu deuten ist, je geringer die Zahl der mit ihm verbundenen und von der Sache her bekannten Tests ist, ist es naheliegend, innerhalb aller

204) Siehe auch VINCENT, D.F. (1953), a.a.O., Seite 108: "It is a maxim in interpreting the results of factor analysis that there must be a reason for each factor that appears, and unless a satisfactory explanation can be found for all or nearly all the factors, the factorisation is considered to have failed".

205) Vgl. Seiten 102-103 und 108-109.

möglichen Faktorstrukturen der als einfache bezeichneten Struktur für erklärende Zwecke den Vorzug zu geben ²⁰⁶⁾.

Die somit angedeutete Trennung zwischen beiden Verfahren durch die verschiedenartige Bedeutung, die den Komponenten und den Faktoren beigemessen wird, ist auf KENDALL zurückzuführen; neuerdings hat sie auch Eingang in den formalen Sprachgebrauch der Statistik gefunden ²⁰⁷⁾. KENDALL faßt die Faktoranalyse als denjenigen Unterfall der Komponentenanalyse auf, in dem eine Interpretation der Faktoren in Übereinstimmung mit einer a priori bestehenden hypothetischen Faktorstruktur versucht wird:

"In component analysis we begin with the observations and look for components in the hope that we may be able to reduce the dimensions of variation and also that our components may, in some cases, be given a physical meaning. We work from the data toward a hypothetical model. In factor analysis we work the other way round; that is to say, we begin with a model and require to see whether it agrees with the data and, if so, to estimate its parameters",

und zur Komponentenanalyse im besonderen ²⁰⁸⁾,

-
- 206) Zu bipolaren Faktoren siehe die Fußnote 177 auf Seite 103. Ganz deutliche Beispiele für Faktoren einer Einfachstruktur (eine hohe und positive Faktorladung, alle weiteren Faktorladungen nahe Null) bietet die Tabelle 9 auf Seite 108.
- 207) Vgl. KENDALL-BUCKLAND (1957), a.a.O., Seite 50-51 (component analysis) und 104-105 (factor analysis). Eine ähnliche Unterscheidung nach der "principal component analysis", "principal factor analysis" (d.h. Anwendung der Methode der Hauptkomponenten auf die reduzierte Korrelationsmatrix) und "factor analysis" stammt von RAO, C.R. (1955) - "Estimation and tests of significance in factor analysis", Psychometrika 20, Seite 93-94. Es muß hinzugefügt werden, daß diese Einteilungen nicht unwidersprochen geblieben sind. Siehe hierzu WHITTLE, P. (1953) - "On principal components and least square methods of factor analysis", Skand. Akt. Tidskr. 35, Seite 223-239. WOLD, H. (1957) - Rezension von KENDALL-BUCKLAND (1957), a.a.O., Kyklos 10, Seite 346. Vgl. auch DE LEVE, G. (1957) - "Enige statistische aspecten van de factoranalyse", Stat. Neerl. 11, Seite 201-209.
- 208) Zum ersten Zitat vgl. KENDALL, M.G. (1957), a.a.O., Seite 37, zum zweiten KENDALL, M.G. (1950), a.a.O., Seite 61.

"We might say that the fundamental problem of component analysis is to economize in the number of variables - a desirable objective whether or not the new variables can be identified with anything 'real'. This is a point which must be decided in each particular case, largely on non-statistical grounds".

Über diese Unterscheidung hinaus geht die Frage, ob die Faktoranalyse tatsächlich ein Forschungsinstrument zur "Entdeckung" von Faktoren in dem Sinne ist, daß durch die Analyse selbst unbekannte Zusammenhänge in der Ausgangsgesamtheit gefunden und damit gleichzeitig als existent nachgewiesen werden können. In letzter Zeit hat sich die Ansicht durchgesetzt, daß es mit Hilfe der Faktoranalyse wohl möglich sein kann, eine bereits vorhandene Hypothese über eine bestimmte Faktorstruktur zu verifizieren, nicht möglich dagegen, von den in einer Analyse aufgefundenen Faktoren ausgehend, die Existenz tatsächlich vorhandener Einflußgrößen als gesichert anzusehen. Den letztgenannten Fall wird man nur dahingehend interpretieren können, daß eine festgestellte Faktorstruktur Ausgangspunkt einer Hypothese sein kann, die erst anschließend mit den der jeweiligen Wissenschaft eigentümlichen Forschungsmethoden gesichert werden muß²⁰⁹⁾. Von besonderer Bedeutung in diesem Zusammenhang ist die Frage, inwieweit Faktoren invariant gegenüber Veränderungen der Ausgangsgesamtheit sind. Dabei werden vor allem für die Einfachstruktur gewisse Ansprüche erhoben; wegen der sehr engen Verbindung dieser Probleme mit fachlichen Fragen der Psychologie, die wir nicht ausreichend beurteilen können, muß auch hier auf die entsprechende

209) Als Beispiel für frühe Ansichten über diesen Gegenstand siehe
DODD, S.C. (1928a) - "The theory of factors, I",
Psychol. Rev. 35, Seite 211-234.
-- Ders. (1928b) - "The theory of factors, II",
Psychol. Rev. 35, Seite 261-279.
Zur gegenwärtigen Auffassung vgl. zum Beispiel
EYSENCK, H.J. (1952) - "Uses and abuses of factor analysis", Appl. Stat. 1, Seite 45-49.

Literatur verwiesen werden ²¹⁰⁾. Eine kurze Zusammenfassung dieser Frage entnehmen wir einer Arbeit von KELLEY ²¹¹⁾:

"There is no search for timeless, spaceless, populationless truth in factor analysis; rather, it represents a simple, straightforward problem of description in several dimensions of a definite group functioning in definite manners, and he who assumes to read moreremote verities into the factorial outcome is certainly doomed to disappointment".

Wenn wir uns nun der Frage der Anwendung der Komponentenanalyse oder der Faktoranalyse außerhalb der Psychologie - damit meinen wir in unserem Falle die Wirtschafts- und Sozialwissenschaften - zuwenden, so ergeben sich als Vorbereitung für den Abschnitt 4 die folgenden Feststellungen.

-
- 210) BARTLETT, M.S. (1937) - "The statistical conception of mental factors", Brit.J.Psychol. 28, Seite 97-104.
McNEMAR, Q. (1942) - "On the number of factors", Psychometrika 7, Seite 9-18.
REYBURN, H.A.; TAYLOR, J.G. (1943) - "On the interpretation of common factors: A criticism and a statement", Psychometrika 8, Seite 53-64.
CATTELL, R.B. (1948) - "The integration of factoranalysis with psychology", J.Educ.Psychol. 39, Seite 227-236.
- - - Ders. (1949) - "A note on factor invariance and the identification of factors", Brit.J.Psychol.Stat. Sect. 2, Seite 134-139.
BABINGTON SMITH, B. (1950) - "A evaluation of factor analysis from the point of view of a psychologist" (Teil 2 von KENDALL, M.G.; BABINGTON SMITH, B. - "Factor analysis"), J.Roy.Stat.Soc.B. 12, Seite 73-85.
THOMSON, G.H. (1951b) - "Factor analysis, its hopes and dangers", Proc. 13th Int. Congr. Psychol. Stockholm, Seite 50-61.
PEEL, E.A. (1953) - Factorial Analysis as a Psychological Technique. In: Uppsala Symposium on Psychological Factor Analysis (Nordisk Psykologi's Monogr. Ser. No. 3), Stockholm: Almqvist & Wiksell, Seite 7-22.
BERNYER, G. (1957) - "Psychological factors: Their number, nature, and identification", Brit.J.Stat.Psychol. 10, Seite 17-27.
HAMILTON, M. (1958) - "An experimental approach to the identification of factors", Brit.J.Stat.Psychol. 11, Seite 159-169.
- 211) KELLEY, T.L. (1940) - "Comment on WILSON and WORCESTER'S 'Note on factor analysis'", Psychometrika 5, Seite 120.

Die Komponentenanalyse als solche, im Abschnitt 1 anwendungsfrei formuliert und im Abschnitt 30 in ihren methodischen Varianten dargestellt, kann als ein exaktes und im Hinblick auf die Erfassung der Variationen optimales Verfahren zur Verringerung der Dimension einer Gesamtheit von Ausgangsvariablen unmittelbar in jedem sachlichen Zusammenhang angewendet werden. Ob ihre Hauptkomponenten interpretiert werden können, ist dabei, wie wir gesehen haben, eine Frage von sekundärer Bedeutung. Daneben ist es möglich, auch im Bereich der Wirtschaftswissenschaften von einem der Faktoranalyse entsprechenden Modell auszugehen. In diesen Fällen jedoch - und das war der Zweck unserer Ausführungen zur Faktoranalyse - wird man sich nach Möglichkeit immer der Methode der Hauptkomponenten, hier als spezielles Verfahren der Faktoranalyse angesehen, zu bedienen haben. Sie kann zwar zur Lösung des Problems der Varianzen h_1^2 der gemeinsamen Faktoren nichts beitragen - diese müssen immer willkürlich bestimmt werden, es sei denn, daß über die Faktorstruktur völlige Gewißheit besteht ²¹²⁾ -, liefert aber nach einer Schätzung der "communalities" das einzige Verfahren, das zu einer in der Methode liegenden und nicht subjektiver Entscheidung (Bewegung der Achsen, Einfachstruktur) überlassenen Lösung führt. Darüber hinaus bietet nach den augenblicklichen Kenntnissen nur die Komponentenanalyse die Gewähr für die einwandfreie Anwendung moderner statistischer Testverfahren ²¹³⁾.

212) Auch die "objektive" Bestimmung der "communalities" nach den auf Seite 110 beschriebenen Verfahren stellt im Sinne der Faktoranalyse eine Willkür dar; denn der kleinste Rang von $R_{(r)}$ ist bei einer als Hypothese vorausgesetzten Faktorstruktur nicht in jedem Falle der "richtige" Rang. Dazu erinnere man sich daran, daß die Zahl, die den Rang der reduzierten Korrelationsmatrix angibt, gleich der Zahl der gemeinsamen Faktoren ist.

213) HOTELLING, H. (1957) - "The relations of newer multivariate statistical methods to factor analysis", Brit. J. Stat. Psychol. 10, Seite 67-79.
KENDALL, M. G. (1957), a.a.O., Seite 93-104.

4 ANWENDUNGEN DER KOMPONENTENANALYSE IN DEN WIRTSCHAFTS-
UND SOZIALWISSENSCHAFTEN

40 ALLGEMEINE ANWENDUNGSPROBLEME; METHODISCHE EINGLIE-
DERUNG DER KOMPONENTENANALYSE

400 Überblick der Anwendungsgebiete

Die enge Verflechtung der Faktoranalyse mit fachlichen Problemen der Psychologie hat bewirkt, daß Anwendungen außerhalb dieser Wissenschaft bisher nur zögernd vorgenommen worden sind. Wir haben in unseren kurzen Ausführungen zur Faktoranalyse bemerkt, wie stark das Ergebnis einer faktoriellen Untersuchung von Entscheidungen, für die es keine methodischen, sondern nur allgemeine oder sachliche Anhaltspunkte gibt, beeinflußt werden kann ²¹⁴⁾. Da die Unterscheidung zwischen Faktoranalyse und Komponentenanalyse noch nicht sehr lange besteht und auch nicht allgemein anerkannt wird, ist auch das von uns im Abschnitt 1 hergeleitete Verfahren von gewissen Hemmungen hinsichtlich der Anwendbarkeit betroffen worden - ganz abgesehen davon, daß die Methode der Hauptkomponenten als spezielles Verfahren zur Lösung eines Ansatzes nach (178) das allgemeine Schicksal der Methoden der

214) Als ein Beispiel für die Einstellung zur Faktoranalyse können wir KENDALL (1950, a.a.O., Seite 60) zitieren, der seine Abhandlung mit den folgenden Sätzen einleitet: "Factor analysis is distinguished from most other branches of statistical technique by having been developed mainly by men who were not primarily statisticians. A few mathematical statisticians have taken an interest in it ... but generally speaking the main developments are due to psychologists. For this reason it has tended to acquire a language and a technique of its own; and, in particular, connected accounts of the subject ... are so bound up with the psychological conception of mental factors that an ordinary statistician has difficulty in seeing it in a proper setting in relation to the general body of statistical method. A feeling even exists among some statisticians that factor analysis is not really an accredited technique at all, that there is something unmentionable about its antecedents which puts it outside respectable statistical usage."

Faktoranalyse teilt. Für die Komponentenanalyse im besonderen gilt außerdem, daß Schwierigkeiten in der rechentechnischen Bewältigung der Methode ihrer Anwendung entgegenstehen können ²¹⁵⁾.

In den nun folgenden Abschnitten wollen wir uns um den Nachweis bemühen, daß es auch in anderen Gebieten als der Psychologie und namentlich in den Wirtschafts- und Sozialwissenschaften durchaus sinnvolle Anwendungen der Komponentenanalyse - sei diese als selbständiges Verfahren oder als Sonderfall der Faktoranalyse betrachtet ²¹⁶⁾ - geben kann. Hierbei werden wir zunächst im Abschnitt 41 eine Reihe von Fällen schildern, in denen die Komponentenanalyse als Forschungsinstrument der Ökonometrie und der Theorie der Indexpzahlen benutzt wurde ²¹⁷⁾. Diese Ausführungen sollen vor allem in den Abschnitten 413 und 414 durch einige Gedanken zu weiteren Anwendungsmöglichkeiten ergänzt werden. Dem Abschnitt 42 bleibt die Besprechung spezieller Verwendungen

215) Vgl. Seite 62-63 und BARGMANN, R. (1956), a.a.O., Seite 12-13.

216) Bei gewissen Anwendungen - zumal solchen, die mehr als zehn Jahre zurückliegen - wird zwischen beiden Formen kein Unterschied gemacht. Wir schließen diese Fälle und auch jene, bei denen für eine Faktoranalyse von einer anderen Methode als der der Hauptkomponenten Gebrauch gemacht wird, ein. Dazu können wir uns auf die ausführliche Erörterung des Zusammenhangs zwischen Komponenten- und Faktoranalyse im Abschnitt 311 stützen und bei den anderen Methoden auf die Diskussion dieser Frage auf Seite 121 verweisen.

217) Eine systematische Zusammenstellung von Beispielen zur Anwendung der Faktoranalyse in den Wirtschafts- und Sozialwissenschaften existiert nach Wissen des Verfassers nicht, vielmehr sind diese recht weit über eine Anzahl von Fachzeitschriften verstreut. Auf eine vollständige Erfassung kann daher nur bedingt Anspruch erhoben werden, zumal eine allem Anschein nach bedeutende Veröffentlichung, nämlich HAMMOND, W.H. (1946) - "Factor analysis as applied to social and economic data", Brit. J. Educ. Psychol. 16, Seite 178, in Deutschland nicht zur Verfügung stand. In der deutschen Literatur konnte als eigentlich wirtschaftsstatistische Anwendung nur die später auf Seite 166 zitierte Arbeit von EGERT ausfindig gemacht werden.

in der Straßenverkehrsunfallstatistik, der statistischen Regionalforschung und der technischen Statistik - ebenfalls mit ergänzenden Anregungen versehen - vorbehalten. Die Klassifizierung der Anwendungsbeispiele ließ sich nicht immer ganz eindeutig durchführen, so kommen besonders Fragen aus dem Bereich der Indexzahlen sowohl in den Abschnitten 412-415 als auch im Abschnitt 421, dort unter regionalstatistischen und soziologischen Aspekten, vor. Bei den auf konkrete Probleme der betrieblichen Praxis, namentlich der Fertigungstechnik, bezogenen Untersuchungen des Abschnitts 422 bleiben alle diejenigen ausgeschlossen, die in das Gebiet der Betriebs- und Arbeitspsychologie gehören. Immerhin sei dazu bemerkt ²¹⁸⁾, daß die Komponentenanalyse als

-
- 218) Untersuchungen dieser Art sind bisher überwiegend für militärische Zwecke vorgenommen worden, jedoch liegen "zivile" Anwendungen ebenso nahe. Vgl. zum Beispiel VERNON, P.E. (1946) - "Statistical methods in the selection of Navy and Army personnel", J.Roy.Stat.Soc. Suppl. 8, Seite 139-153. CARTER, L.F.; DUDEK, F.J. (1947) - "The use of psychological techniques in measuring and critically analyzing navigators' flight performance", Psychometrika 12, Seite 31-42. TAYLOR, C.W. (1947) - "A factorial study of fluency in writing", Psychometrika 12, Seite 239-262. BARAKAT, M.K. (1951) - "A factorial study of mathematical abilities", Brit.J.Psychol.Stat.Sect. 4, Seite 137-156. GUILFORD, J.P. (1952) - "Factor analysis of the Army Air Forces SHEPPARD field battery of experimental aptitude tests", Psychometrika 17, Seite 45-63. SHARMA, O.C. (1954) - "Factor analysis of technical trades and educational examination marks of the air-craftsmen of the Indian Air Force", Sankhya 13, Seite 27-34. GLICKMAN, A.S. (1958) - "Factor analysis of personnel components of ship performance", Opn.Res. 6, Seite 106-115. Zur Rolle der Psychologie in der Berufsberatung siehe den Abschnitt "Statistiques et orientation professionnelle" bei TECOZ, H.F. (1944) - "Utilité des méthodes statistiques dans quelques applications de la psychologie (Données nouvelles)", Zeitschr.f.Schweiz.Stat.u. Volksw. 80, Seite 48-77.

Bestandteil faktoranalytischer Forschungen dieser Art auch für wirtschaftliche Belange recht bedeutsam ist; dabei ist etwa an die Konstruktion von Eignungstests für bestimmte technische und administrative Aufgaben in Betrieben oder für allgemeine Probleme der Berufsberatung und -lenkung gedacht. Auch auf soziologische Arbeiten gehen wir nur insoweit ein, als Fragestellungen der Soziologie in den aus anderen Gründen behandelten Veröffentlichungen von HAGOOD und PRICE berührt werden. Wir wollen jedoch erwähnen, daß die Faktoranalyse darüber hinaus in mehreren Fällen rein soziologischen und auch politischen Forschungsaufgaben dienstbar gemacht worden ist. Hierzu gehören Studien über das Verhalten von Wählern und über die Faktoren der politischen sowie religiösen Meinungsbildung, ferner, um nur ein weiteres Beispiel ²¹⁹⁾ zu nennen, die Untersuchung des Einflusses sozialer Bedingungen auf die Sterblichkeit.

401 Die Stellung der Komponentenanalyse im Rahmen anderer statistischer Verfahren

Bevor auf konkrete Fälle der Anwendung eingegangen wird, sei ein von KENDALL aufgestelltes Schema ²²⁰⁾ betrachtet,

219) Die folgende Zusammenstellung bietet eine Auswahl entsprechender Veröffentlichungen:

GOSNELL, H.F.; SCHMIDT, M.J. (1936) - "Factorial and correlational analysis of the 1934 vote in Chicago", J. Am. Stat. Ass. 31, Seite 507-518.

KLINGBERG, F.L. (1941) - "Studies in measurement of the relations among sovereign states", Psychometrika 6, Seite 335-352.

BUCKATZSCH, E.J. (1947) - "The influence of social conditions on mortality rates", Pop. Stud. 1, Seite 229-248.

SANAI, M. (1952) - "An empirical study of political, religious, and social attitudes", Brit. J. Psychol. Stat. Sect. 5, Seite 81-91.

SCHUESSLER, K.F.; DRIVER, H. (1956) - "A factor analysis of sixteen primitive societies", Am. Soc. Rev. 21, Seite 493-499.

LOCKE, H.J.; WILLIAMSON, R.C. (1958) - "Marital adjustment: A factor analysis study", Am. Soc. Rev. 23, Seite 562-569.

220) KENDALL, M.G. (1950), a.a.O., Seite 60-61.

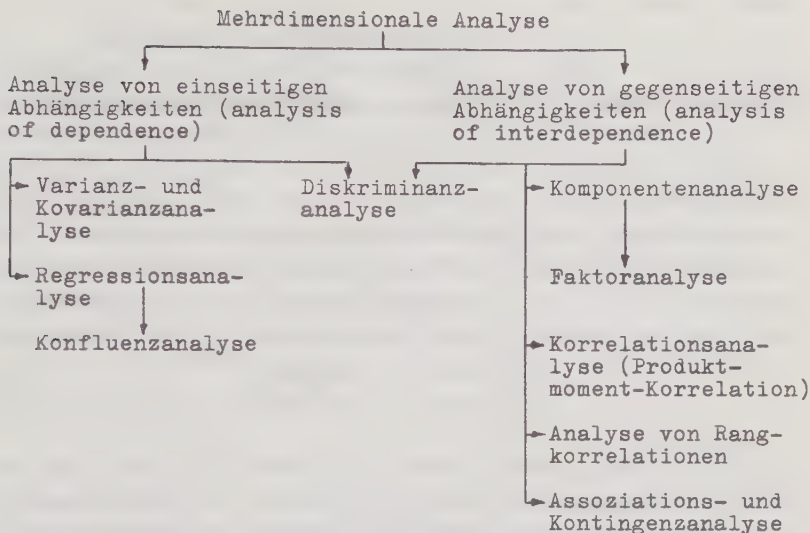
durch das die Positionen der Komponenten- und Faktoranalyse innerhalb eines größeren Kreises statistischer Methoden charakterisiert werden. Hierzu nimmt KENDALL in der mehrdimensionalen Analyse eine grundsätzliche Trennung zwischen Verfahren, mit denen im voraus festgelegte, einseitige Abhängigkeiten von Variablen untersucht werden (analysis of dependence) und solchen vor, durch welche die gegenseitigen, strukturellen Abhängigkeitsbeziehungen zwischen den Variablen zu ermitteln sind (analysis of interdependence) ²²¹):

"I find it useful to draw a distinction between the analysis of dependence and the analysis of interdependence. In the latter we are interested in how a group of variates are related among themselves, no one being marked out by the conditions of the problem as of greater prior importance than the others, whereas in the analysis of dependence we are interested in how a certain specified group (the dependent variates) depend on the others."

Die Komponentenanalyse hat unter diesen Voraussetzungen ihren Platz in der zweiten Gruppe, da unter den Variablen x_i ($i=1,2,\dots,n$) der Ausgangsgesamtheit keine Abhängigkeitsbeziehungen postuliert werden und Abhängigkeiten letzten Endes nur zwischen den x_i und den gesuchten, nicht in der Ausgangsgesamtheit enthaltenen Komponenten z_k ($k=1,2,\dots,p$) bestehen ²²²). Das Schema von KENDALL hat folgendes Aussehen:

221) KENDALL, M.G. (1950), a.a.O., Seite 60.

222) Dabei ist es nicht ausgeschlossen, daß einige der z_k so hohe Korrelationen mit einer Reihe der x_i aufweisen, daß sie praktisch mit diesen gleichgesetzt und daher die x_i in abhängige und unabhängige Variablen eingeteilt werden können. Jedoch wäre eine solche Lösung das Resultat, nicht der Ausgangspunkt einer Analyse.



Die in dieser Anordnung enthaltenen Stellungen der einzelnen Verfahren zueinander sind mehrfach Gegenstand spezieller Arbeiten gewesen. Wenn man als Zielsetzung der Varianzanalyse ²²³⁾ die Zerlegung der Variation einer Gesamtheit von Variablen nach Ursachengruppen für bestimmte, in der Regel parametrische Verfahren zur Prüfung von Hypothesen bezeichnen will, so besteht dazu in der Komponentenanalyse eine Parallele in der schrittweisen Erfassung der Varianzen der x_i durch die Komponenten z_k . Diese Beziehungen sind von BURT und CREASY untersucht worden ²²⁴⁾. Weiterhin existieren Zusammenhänge zwischen der Komponentenanalyse und

223) Siehe etwa
YULE, G.U.; KENDALL, M.G. (1950¹⁴) - An Introduction to the Theory of Statistics. London: Charles Griffin & Co., Seite 503-529.
HOEL, P.G. (1957), a.a.O., Seite 248-264.
KENDALL-BUCKLAND (1957), a.a.O., Seite 311-312.

224) BURT, C. (1948a) - "A comparison of factor analysis and analysis of variance", Brit. J. Psychol. Stat. Sect. 1, Seite 3-26.
CREASY, M.A. (1957) - "Analysis of variance as an alternative to factor analysis", J. Roy. Stat. Soc. B. 19, Seite 318-325.

verschiedenen Problemen der multiplen und partiellen Korrelation sowie der Regressionsanalyse. Dabei hat sich vor allem gezeigt, daß der formale Ansatzpunkt der Komponentenanalyse - nämlich die Korrelationsmatrix $\mathcal{R}$, gegebenenfalls auch die Matrix $\mathcal{R}_{(r)}$ ²²⁵⁾ - sich sehr gut für die vereinfachte Berechnung multipler oder partieller Korrelationskoeffizienten eignet, damit zugleich für die Bestimmung der Parameter von komplizierten Regressionsgleichungen ²²⁶⁾. Schließlich ist auch die Fragestellung der kanonischen Korrelationen ²²⁷⁾ - die größtmögliche Korrelation zwischen den linearen Kombinationen mehrerer Gruppen von Variablen - mit der der Komponentenanalyse verwandt ²²⁸⁾.

225) Vgl. (36) auf Seite 23 und (36''') auf Seite 108.

226) DWYER, P.S. (1940) - "The evaluation of multiple and partial correlation coefficients from the factorial matrix", Psychometrika 5, Seite 211-232.

WOOLF, B. (1951) - "Computation and interpretation of multiple regressions", J. Roy. Stat. Soc. B. 13, Seite 100-119, insbesondere Seite 117.

GUTTMAN, L. (1957) - "Simple proofs of relations between the communality problem and multiple correlation", Psychometrika 22, Seite 147-157.

CREAGER, J.A. (1958) - "General resolution of correlation matrices into components and its utilization in multiple and partial regression", Psychometrika 23, Seite 1-8.

227) HOTELLING, H. (1936b) - "Relations between two sets of variates", Biometrika 28, Seite 321-377.

TINTNER, G. (1952), a.a.O., Seite 114-121.

KENDALL, M.G. (1957), a.a.O., Seite 68-85.

HOOPER, J.W. (1959) - "Simultaneous equations and canonical correlation theory", Econometrica 27, Seite 245-256.

228) THOMSON, G.H. (1947) - "The maximum correlation of two weighted batteries (HOTELLING'S 'Most predictable criterion')", Brit. J. Psychol. Stat. Sect. 1, Seite 27-34.

BURT, C. (1948b) - "Factor analysis and canonical correlations", Brit. J. Psychol. Stat. Sect. 1, Seite 95-106.

KENDALL, M.G. (1957), a.a.O., Abschnitt 5.12 auf Seite 81.

41 ÖKONOMETRIE UND INDEXZAHLEN

410 Die Hauptkomponenten volkswirtschaftlicher Bewegungsvorgänge: Eine Untersuchung von STONE

Eine bedeutsame Anwendung der Komponentenanalyse in der Ökonometrie ist im Jahre 1947 von STONE ²²⁹⁾ veröffentlicht worden. Obwohl sie nicht der einzige Fall ist, gilt sie wegen einer Reihe von bemerkenswerten Ergebnissen als gleichsam "klassisches" Anwendungsbeispiel ²³⁰⁾. Wir wollen diese Untersuchung im folgenden näher behandeln.

STONE geht von 17 makroökonomischen Größen (blocks of transactions) aus, die in den USA durch den Zeitraum von 1922 bis 1938 beobachtet wurden ²³¹⁾. Im Sinne unserer Terminologie sind die Zeitreihen als Variable ($n=17$), die einzelnen Jahre als Beobachtungspunkte ($N=17$) aufzufassen. Es folgt eine Zusammenstellung der Variablen. ²³²⁾

229) STONE, R. (1947) - "On the interdependence of blocks of transactions", J. Roy. Stat. Soc. Suppl. 9, Seite 1-32.

230) TINTNER (1952, a.a.O., Seite 107-109) und KENDALL (1957, a.a.O., Seite 26-27) zitieren STONE in einiger Ausführlichkeit.

231) Die Daten stammen aus:
KUZNETS, S. (1941) - National Income and its Composition, 1919-1938. New York: National Bureau of Economic Research.
BARGER, H. (1942) - Outlay and Income in the United States 1921-1938, Seiten 42-43, 50-51, 58-59 und 62-63. (Angaben nach STONE).

232) Vgl. STONE, R. (1947), a.a.O., Table I (Components of Income and Outlay in the United States of America 1922-1938) auf Seite 11. Die Variablen 2 und 4 wurden von STONE zusammengefaßt, weil ihre Bestandteile eine Korrelation von mehr als 0,95 aufwiesen.

Tabelle 10: Bezeichnungen von 17 makroökonomischen Variablen nach KUZNETS und BARGER

Variable	Einkommen (E) Ausgaben (A)	Bezeichnung der Variablen
(1)	(2)	(3)
1	E	Bruttolohnsumme (Employees' Compensation)
2	A	Konsumgüter + Investitions- güter (Consumers' perishable Goods plus Producers' durable Goods)
3	E	Nettoersparnis der Unterneh- men + Kapitalverzinsung (Unadjusted Net Savings of Enterprises plus Adjustment for Capital Revaluation)
4	A	Quasi-langlebige + langlebige Konsumgüter (Consumers' semi- durable Goods plus Consumers' durable Goods)
5	A	Dienstleistungen an Private (Consumers' Services)
6	A	Bautätigkeit (Construction)
7	A	Nettoausgaben des Staates (Net Public Outlay)
8	A	Nettozunahme der Lagerhaltung (Net Increase in Inventories)
9	E	Verzinsung der Lagerbestände (Inventory Revaluation Ad- justment)
10	E	Nettomieten an Private (Net Rent received by Individuals)
11	E	Privatentnahmen der Unter- nehmer (Entrepreneurial With- drawals)
12	E	Verteilte Dividenden (Dividends)
13	-A	Abschreibungen (Adjustment for Depreciation)
14	E	Zinsen (Interest)
15	E	Verteilte Dividenden, Zinsen und nicht-gewerbliche Zahlungen aus dem Ausland - direkte Steuern von Privaten + Ent- schädigungen für Kriegsteil- nehmer + Leistungen der

Fortsetzung der Tabelle 10:

Variable	Einkommen (E) Ausgaben (A)	Bezeichnung der Variablen
(1)	(2)	(3)
15 (fortgesetzt)	E	Sozialversicherung - Sozialversicherungsbeiträgen von Arbeitnehmern (Dividends, Interest and non-commercial Remittances from Abroad less direct Taxes paid by Individuals plus Veterans' Bonus plus Social Security Benefits less Employees' Social Security Contributions)
16	-E	Abschreibungen (Adjustment for Depreciation and Depletion)
17	A	Zahlungsbilanz einschließlich Ausgaben von Ausländern in den USA (Foreign Balance including foreign Tourist Expenditure in the United States)

Die zu jeder dieser 17 Variablen gegebenen 17 Werte wurden auf einen Mittelwert 0 bezogen; da sie alle in Geldeinheiten ausgedrückt sind (1 Einheit = 1 Million US-Dollar), war eine vollständige Normierung nach (4) nicht notwendig. Als Ziel der Komponentenanalyse gibt STONE an ²³³⁾:

"First, how far can the variance of a number of transaction blocks be satisfactorily explained in terms of a small number of factors ?

Second, if it is found that most of the total variance can be explained by a few factors, how far does this satisfactory result extend to the individual variances of the observed variables ?

Third, is it possible to identify the factors with series estimated from economic considerations and not from the factor-analysis itself ?"

233) STONE, R. (1947), a.a.O., Seite 11.

Die erste dieser Fragen gibt das von uns als Hauptanliegen der Komponentenanalyse bezeichnete Problem der Verringerung der Dimension der Ausgangsgesamtheit wieder, die zweite und die dritte betreffen zusätzliche Erkenntnisse aus der Modellvorstellung der Faktoranalyse, die jedoch - wie wir im Abschnitt 311 gesehen hatten - durchaus mit als Zielsetzung der Komponentenanalyse betrachtet werden können. Mit der dritten Frage wollen wir uns erst im Abschnitt 411 beschäftigen.

Methodisch geht STONE nach HOTELLING vor, allerdings wird an Stelle der Korrelationsmatrix die der Summen der Abweichungsquadrate und -produkte ²³⁴⁾ benutzt. Diese 17-reihig quadratische, symmetrische Matrix liefert 16 reelle und positive Eigenwerte ²³⁵⁾ ($p=16$), die im einzelnen nicht alle angegeben werden. Das ist deshalb nicht notwendig, weil die ersten drei Hauptkomponenten 97,45% der Gesamtvariation in der Ausgangsgesamtheit erfassen. Darüber gibt die folgende Übersicht, die sich an der Tabelle 3 orientiert, Aufschluß ²³⁶⁾.

234) Dies entspricht der Formulierung von VAN DEN BOGAARD-KAPTEIN, jedoch normieren diese die Summen der Abweichungsquadrate, die "Devianzen", auf 1. Das gilt hier nur für die Komponenten. Vgl. Seite 96-97.

235) Die Matrix befindet sich bei STONE (1947, a.a.O.) als Table II (Sums of Squares and Products of Deviations from Means) auf Seite 12. Sie hat den Rang 16, also einen verschwindenden Eigenwert. Das wird von STONE in der Fußnote auf Seite 16 damit erklärt, daß wegen der kleinen Zahl ($N=17$) die Abweichungen vom arithmetischen Mittel in der der Matrix ~~36~~ entsprechenden Matrix nur 16 linear unabhängige Zeilen (16 Freiheitsgrade !) existieren können. Dieser Rang überträgt sich in Analogie zu (30) auf die Matrix der Summen der Abweichungsquadrate und -produkte.

236) STONE, R. (1947), a.a.O., Table IV auf Seite 15. Die Differenz von 0,01% dürfte durch Rundungen entstanden sein. - Unsere Tabelle 3 befindet sich auf Seite 77, die weiter unten genannten Tabellen 5 und 6 auf Seite 81.

Tabelle 11: Anteilswerte der ersten drei Hauptkomponenten an der Gesamtvariation der 17 Ausgangsvariablen von STONE

Eigenwert	Varianzanteile (in v.H.)	
	je Eigenwert	kumuliert
(1)	(2)	(3)
$\lambda_1 = 1605,3132$	0,8076	0,8076
$\lambda_2 = 210,5318$	0,1059	0,9135
$\lambda_3 = 121,1148$	0,0609	0,9744
Summe	0,9744	

Im Hinblick auf die zweite Problemstellung, nämlich die Aufgliederung der Gesamtvarianz in die Varianzen der einzelnen Ausgangsvariablen und deren "explanation" durch die Hauptkomponenten im Sinne der Tabellen 5 und 6, ergibt sich nachstehendes Bild ²³⁷⁾:

Tabelle 12: Anteilswerte der ersten drei Hauptkomponenten an den Varianzen der 17 Ausgangsvariablen von STONE

Ausgangsvariable	Varianzanteile der Hauptkomponenten (in v.H.)			
	z_1	z_2	z_3	z_4 bis z_{16}
(1)	(2)	(3)	(4)	(5)
1	0,9666	0,0229	0,0089	0,0016
2	0,9068	0,0009	0,0851	0,0072
3	0,5018	0,4918	0,0006	0,0058
4	0,9462	0,0087	0,0296	0,0155
5	0,8267	0,1215	0,0431	0,0087
6	0,7311	0,0051	0,2581	0,0057
7	0,4992	0,0019	0,3367	0,1622
8	0,4230	0,1010	0,1093	0,3667
9	0,0889	0,7209	0,0252	0,1650
10	0,5467	0,0332	0,3961	0,0240
11	0,8430	0,0558	0,0849	0,0163
12	0,7194	0,1382	0,0590	0,0834
13	0,8270	0,1333	0,0066	0,0331
14	0,0000	0,7486	0,0001	0,2513
15	0,2108	0,0001	0,0537	0,7354
16	0,7078	0,0085	0,0074	0,1963
17	0,0447	0,0051	0,3883	0,5619

237) STONE, R. (1947), a.a.O., Table X auf Seite 21. Vgl. auch TINTNER, G. (1952), a.a.O., Seite 108.

Diese Aufstellung zeigt, daß das für die Gesamtvariation (Tabelle 11) hinsichtlich der Repräsentation festgestellte Ausreichen der ersten drei Hauptkomponenten im wesentlichen auch für jede einzelne der Ausgangsvariablen x_i' Gültigkeit besitzt. Das geht am deutlichsten aus der Spalte 5 hervor, in der sich die in diesem Falle den Charakter von Residuen tragenden Werte

$$(184) \quad \text{Res Var } x_i = \text{Var } x_i - \sum_{k=1}^3 a_{ik}^2 = \sum_{k=4}^{16} a_{ik}^2$$

($i=1,2,\dots,17$),

in unseren Symbolen dargestellt, befinden ²³⁸⁾. Wie man sieht, sind die Hauptkomponenten z_1 , z_2 und z_3 nur für die Wiedergabe der Variablen 8, 14, 15 und 17 unzureichend ²³⁹⁾. Insgesamt jedoch darf man aus den Tabellen 11 und 12 den Schluß ziehen, daß zur Bildung des Systems (11) die ersten drei Hauptkomponenten für eine gute Annäherung genügen werden oder daß, im Sinne von (111) ausgedrückt, die Bewegung der x_i' ($i=1,2,\dots,17$) im Zeitablauf mit befriedigender Genauigkeit durch die der z_1 , z_2 und z_3 dargestellt werden kann. Wir deuten beide Gesichtspunkte in den Tabellen 13 und 14 an ²⁴⁰⁾.

238) Vgl.(38), Seite 24, und (86), Seite 45. Tatsächlich wären die Zahlen der Tabelle 12 wegen der bei STONE nicht vollständigen Normierung der Variablen formelmäßig etwas anders auszudrücken.

239) Zur sachlichen Begründung des Verhaltens der Hauptkomponenten in Bezug auf die einzelnen Variablen siehe STONE,R.(1947), a.a.O., Seite 22-30.

240) STONE,R.(1947), a.a.O., Tables IV und VI, Seite 15 -16. Die Werte der Tabelle 13, Seite 135, entsprechen unter Nichtbeachtung der Normierung den Elementen a_{ik} der Matrix A , sie liefern die gegenüber (11) unvollständigen Schätz-(Regressions-)gleichungen, zum Beispiel $x_i' = 24,9622z_1 - 3,8396z_2 - 2,3958z_3 \dots$. Vgl. auch Tabelle 9 auf Seite 108. Die Tabelle 14, Seite 135, entspricht der Matrix $\hat{A}$.

Tabelle 13: Matrix der den Größen a_{ik} entsprechenden Koeffizienten (nicht normierte Faktoraladungen) nach STONE (Auszug)

Variable	Koeffizienten der Hauptkomponenten		
	z_1	z_2	z_3
(1)	(2)	(3)	(4)
1	24,9622	-3,8396	-2,3958
2	17,2581	-0,5389	-5,2854
⋮	⋮	⋮	⋮

Tabelle 14: Matrix der Hauptkomponenten nach STONE (Auszug)

Jahr	Hauptkomponente		
	z_1	z_2	z_3
(1)	(2)	(3)	(4)
1922	-0,1316	0,3258	0,2737
1923	0,0800	0,2837	0,1707
⋮	⋮	⋮	⋮

Damit haben wir die Ergebnisse von STONE unter dem Gesichtspunkt der Verringerung der Dimension einer Ausgangsgesamtheit in einiger Ausführlichkeit betrachtet. Die Bedeutung dieser "reduction in effective dimension-number" (KENDALL) für ökonometrische Analysen wird im vorliegenden Fall wohl am besten von TINTNER mit den Worten ²⁴¹⁾

"It is apparent that we can represent the seventeen variables ... reasonably well by not more than three factors. Together they account for more than 97 per cent of the total variance of all variables. Hence in a statistical analysis of economic data the use of three principal components instead of seventeen original variables will save 14 degrees of freedom. This is conceivably an important gain because of the short length of many economic series available for statistical analysis "

gekennzeichnet. Diese Aussage erhält ein ganz besonderes Gewicht dann, wenn für die Beobachtung ökonomischer

241) TINTNER, G. (1952), a.a.O., Seite 109.

Bewegungsabläufe großen Ausmaßes, etwa bei der Aufstellung eines ökonomischen Modells für eine gesamte Volkswirtschaft²⁴²⁾, die aus theoretischen Gründen einzubeziehende Zahl der Variablen die für die üblichen multiplen Regressionsansätze gegebenen numerischen Möglichkeiten übersteigt; hier kann die Komponentenanalyse sowohl bei gegebenem Modell den Arbeitsumfang verringern als auch bei feststehender Rechenkapazität die Erweiterung des Modells durch Einbeziehung zusätzlicher, jedoch nur durch Hauptkomponenten vertretener Variablen ermöglichen. Diese Fragen sind von STONE und VAN EIJK²⁴³⁾ angedeutet und in neuester Zeit von THEIL, KLOEK sowie MENNES²⁴⁴⁾ hinsichtlich der von TINTNER erwähnten Problematik der durch die Kürze von Zeitreihen verursachten Restriktionen eingehend untersucht worden. Es besteht außerdem ein enger Zusammenhang zwischen der Komponentenanalyse und der von DAVIS²⁴⁵⁾ zur Feststellung der Zahl der Freiheitsgrade ökonomischer Zeitreihen entwickelten "method of elementary energies".

242) Vgl. zum Beispiel

KLEIN, L.R.; GOLDBERGER, A.S. (1955) - An Econometric Model of the United States 1929-1952 (Contributions to Economic Analysis IX). Amsterdam: North-Holland Publishing Co., ferner das mit 72 Variablen ausgestattete ökonomische Modell für England, welches vom Oxford University Institute of Statistics für vierteljährliche Prognosen benutzt wird. Siehe BALL, R.J.; HAZLEWOOD, A.; KLEIN, L.R. (1959) - "Econometric forecasts for 1959", Bull. Oxf. Univ. Inst. Stat. 21, Seite 3-16.

243) STONE, R. (1947), a.a.O., Seite 32.

EIJK, C.J. VAN (1957) - "Economische toepassingen van factoranalyse", Stat. Neerl. 11, Seite 211-213.

244) THEIL, H.; KLOEK, T. (1959) - "The statistics of systems of simultaneous economic relationships", Stat. Neerl. 13, Seite 76-78.

KLOEK, T.; MENNES, L.B.M. (1960) - "Simultaneous equations estimation based on principal components of predetermined variables", Econometrica 28, Seite 45-61.

245) DAVIS, H.T. (1941) - The Analysis of Economic Time Series (COWLES Commission for Research in Economics Monogr. No. 6). Bloomington: The Principia Press, Seite 198-207.

411 Identifizierung der Hauptkomponenten mit ökonomischen Variablen

Die Untersuchung von STONE ist neben ihrer Bedeutung als Anwendungsfall der Komponentenanalyse durch das Gelingen einer plausiblen Identifizierung ihrer drei Hauptkomponenten im Sinne der Faktoranalyse ausgezeichnet. Hierzu wollen wir an unsere Bemerkungen zur Unterscheidung zwischen Hauptkomponenten und Faktoren im Abschnitt 311 anknüpfen und zunächst feststellen, daß die in der vorliegenden Analyse aufgefundenen Hauptkomponenten z_1 , z_2 und z_3 zwar im Hinblick auf die von ihnen geforderten algebraischen Eigenschaften optimale, im übrigen aber sachlich beziehungslose Größen sind. Ihre "Deutung" aus dem Wesen der Methode heraus ist nicht möglich ²⁴⁶⁾, jedes Bemühen um eine Identifizierung muß sich an der Sache orientieren.

Für die hier diskutierte Untersuchung führen diese Überlegungen zu der Einstellung, die Reduktion der Ausgangsgesamtheit auf drei Variable als entscheidend anzusehen und nicht den Versuch zu unternehmen, jede der drei Hauptkomponenten einzeln mit realen ökonomischen Variablen in Verbindung zu bringen. Läßt sich ein System von drei ökonomischen Variablen finden, das aus theoretischen Gründen für eine Erklärung der Ausgangsvariablen hinreichend ist, so kann dessen Zusammenhang mit dem in derselben Dimension befindlichen System der Hauptkomponenten geprüft werden - unter Umständen können beide Systeme durch eine Bewegung der Komponentenachsen ineinander überführt werden ²⁴⁷⁾. STONE gibt an, daß vor Durchführung der Komponentenanalyse die Vermutung bestand, die Variation in den 17 Ausgangsvariablen ließe sich weitgehend durch die drei Reihen

246) Die Gründe dafür haben wir, wie bereits erwähnt, im Abschnitt 311, Seite 115-121, ausführlich dargelegt. Vgl. auch STONE, R. (1947), a.a.O., Seiten 19 und 30.

247) Das entspräche der auf den Seiten 112-113 behandelten Veränderung der Faktorstruktur durch orthogonale bzw. schiefwinklige Drehung der Achsen.

Volkseinkommen Y (total income or some similar quantity), Veränderungsrate des Volkseinkommens ΔY (rate of change of total income) und einer die Expansions- oder Kontraktionstendenz der Volkswirtschaft ausdrückenden - hier linear angenommenen - Trendgröße t erklären. Das bemerkenswerte Endergebnis der Analyse besteht nun in der guten Übereinstimmung, die zwischen den berechneten Hauptkomponenten und den drei genannten Variablen existiert. Die Korrelationskoeffizienten für alle sechs Reihen zeigt die folgende Aufstellung 248).

Tabelle 15: Korrelationskoeffizienten zwischen Hauptkomponenten und ökonomischen Variablen nach STONE

	z_1	z_2	z_3	Y	ΔY	t
z_1	1	-	-	-	-	-
z_2	0	1	-	-	-	-
z_3	0	0	1	-	-	-
Y	<u>0,995</u>	-0,041	0,057	1	-	-
ΔY	-0,056	<u>0,948</u>	-0,124	-0,102	1	-
t	-0,369	-0,282	<u>-0,836</u>	-0,414	-0,112	1

Die in dieser Tabelle ausgewiesenen Zahlenwerte für die Stärke des Zusammenhangs zwischen den Variablen Y und z_1 , ΔY und z_2 sowie t und z_3 berechtigen zu der Aussage, daß die Gesamtvariation der Ausgangsgesamtheit in guter Annäherung durch ein dreidimensionales System von ökonomischen Variablen dargestellt werden kann, das wegen seiner Übereinstimmung mit den Hauptkomponenten in Bezug auf die Verringerung der Dimension optimal ist. In den durch die Eintragungen der Tabelle 12 gezogenen Grenzen kann diese Aussage auf die meisten der einzelnen Ausgangsvariablen

248) STONE, R. (1947), a.a.O., Table IX auf Seite 20. Wegen der Symmetrie der Tabelle sind die Koeffizienten nur einmal aufgeführt. Die sehr hohe Korrelation zwischen z_1 und Y erscheint erklärlich dadurch, daß das Volkseinkommen eine aus einem Teil der Ausgangsvariablen zusammengesetzte Größe ist. Dagegen wurde ΔY von der Ausgangsgesamtheit unabhängig aus Vierteljahreswerten berechnet, über die Größe t fehlen nähere Angaben.

erweitert werden, so daß sich als Näherungsausdruck die allgemeine Beziehung 249)

$$(185) \quad x = aY + b \frac{dY}{dt} + ct$$

aufstellen läßt.

Wir wollen unsere Ausführungen über die Untersuchung von STONE nicht ohne einen Hinweis auf gewisse Mängel abschließen. Hierzu sind in erster Linie die Gefahren zu nennen, die aus dem Zeitreihencharakter der Ausgangsvariablen erwachsen. Das gilt nicht nur für die dadurch bedingte geringe Anzahl der für die Variablen gegebenen Werte - N ist gleich n , eine bei der Analyse von Zeitreihen sehr häufig hinderliche Erscheinung 250) - , sondern auch für die völlig unberücksichtigt gebliebene Frage der Verzerrung durch Auto-Korrelation 251). Schließlich sind die Ergebnisse der mit einem beträchtlichen Rechenaufwand (Eigenwertbestimmung einer 17-reihigen Matrix) belasteten Analyse keineswegs ungewöhnlich, denn sie bestätigen nur eine aus allgemeinen theoretischen Erwägungen abgeleitete und als solche genügend begründete Vermutung über das Bestehen gewisser Zusammenhänge in der Ausgangsgesamtheit. Jedoch möchten wir in diesem Umstand keine entscheidende Einwendung sehen; es gibt zahlreiche Möglichkeiten einer entsprechenden Verwendung der Komponentenanalyse in der Ökonometrie - etwa bei

249) STONE, R. (1947), a.a.O., Gleichung (10) auf Seite 21. (185) ist lediglich eine Verallgemeinerung des in der Fußnote 240, Seite 134, angedeuteten Gleichungssystems.

250) Zur Rolle der Komponentenanalyse als Methode zur "Vergrößerung" der Differenz zwischen N und n durch Substitution von n durch p ($p < n$), also zur Frage der Freiheitsgrade, vgl. Seite 135-136.

251) Siehe TINTNER, G. (1940) - "The analysis of economic time series", J. Am. Stat. Ass. 35, Seite 93-100, DAVIS, H. T. (1941), a.a.O., Seite 102-141, TINTNER, G. (1952), a.a.O., Seite 240-252 und, speziell zur Problematik der Korrelation von Zeitreihen, ANDERSON, O. (1957), a.a.O., Seite 212-218.

Nachfrageanalysen ²⁵²⁾ - bei denen die Zusammenhänge nicht ohne weiteres sichtbar sind und für die die Arbeit von STONE als einführendes Musterbeispiel großen Wert besitzt.

412 Einige Anwendungen auf Indexzahlen

Ein weiteres Anwendungsgebiet der Komponentenanalyse ist das der Indexzahlen, wobei neben einigen speziellen Problemen sowohl die Konstruktion einzelner Indizes wie auch die Zusammenfassung von Gruppenindizes zu einem Gesamtindex als Ansatzpunkte in Erwägung zu ziehen sind. Wir werden diese Fragen in den Abschnitten 413 und 414 genauer untersuchen und wollen uns zunächst mit einigen Beispielen aus der Literatur beschäftigen.

Die erste Anwendung dieser Art scheint eine Arbeit von RHODES zu sein ²⁵³⁾, in der versucht wird, den Gedanken der Faktorstruktur auf wirtschaftliche Tatbestände zu übertragen und für die 14 Indexreihen ($n=14$), die dem von der

252) Vgl. die diesbezüglichen Anregungen bei STONE, R. (1945) - "The analysis of market demand", J. Roy. Stat. Soc. 108, Seite 286-382, STONE, R. (1947), a.a.O., Seiten 30 und 32, sowie als Einzelbeispiele LANSING, J.B.; BLOOD, D.M. (1958) - "A cross-section analysis of non-business air travel", J. Am. Stat. Ass. 53, Seite 928-947. JANOSI, P.E. DE (1959) - "Factors influencing the demand for new automobiles", J. Marketing 23, Seite 412-418. Eine ganze Reihe weiterer Anregungen, die wir hier nicht näher betrachten können, enthält der mit einer ausführlichen Bibliographie versehene Abschnitt "Types of Problems to which Correlation and Regression Analysis have been applied" (Kapitel 25) bei EZEKIEL, M.; FOX, K.A. (1959) - Methods of Correlation and Regression Analysis. New York: J. Wiley & Sons (London: Chapman & Hall), Seite 434-467. Vgl. ferner Abschnitt 414, Seite 159-160.

253) RHODES, E.C. (1937) - "The construction of an Index of Business Activity", J. Roy. Stat. Soc. 100, Seite 18-39.

Zeitschrift "The Economist" seit 1933 für England berechneten "Index of Business Activity" zugrunde liegen ²⁵⁴⁾, einen gemeinsamen Faktor "business activity" zu bestimmen. RHODES benutzt zwar die Modellbildung der Faktoranalyse, verwendet jedoch keine Hauptkomponenten; wir gehen deshalb auf die Einzelheiten seiner Methode nicht ein. Immerhin sei bemerkt, daß die Faktorladungen ein System von Gewichten für die Zusammensetzung der einzelnen Indizes ergaben, das mit dem ursprünglichen Schema des "Economist"-Index gut zu vereinbaren war ²⁵⁵⁾. Im Gegensatz zu RHODES haben KENDALL und TINTNER für die Konstruktion von Indizes die Komponentenanalyse herangezogen.

KENDALL ²⁵⁶⁾ untersuchte die vom englischen Landwirtschaftsministerium veröffentlichten Ernteergebnisse für 10 Fruchtarten - Weizen, Gerste, Hafer, Bohnen, Erbsen, Kartoffeln, Futterrüben (einschl. schwedischer Kohlrüben), Runkelrüben, Futterpflanzen (Klee, Luzerne, Esparsette, englisches Raygras und andere) sowie Heu - in 48 Bezirken Englands (ohne Schottland und Wales). Der Grundgedanke der Analyse bestand darin, ein Maß für die allgemeine Produktivität der Landwirtschaft ²⁵⁷⁾ - bezogen auf die angegebenen Fruchtarten,

254) CROWTHER, G. (1934) - "The 'Economist' Index of Business Activity", J. Roy. Stat. Soc. 97, Seite 241-261.

255) RHODES, E. C. (1937), a. a. O., Seite 37-39.

256) KENDALL, M. G. (1939) - "The geographical distribution of crop productivity in England", J. Roy. Stat. Soc. 102, Seite 21-48.

- - - Ders. (1957), a. a. O., Seite 18-19.

257) "Produktivität" ist hier nicht mit "Bodenfruchtbarkeit" gleichzusetzen, sondern als erweiterter Begriff aufzufassen. KENDALL (1939, a. a. O., Seite 24) schreibt: "'Productivity' is not a synonym for 'fertility'! I use the word to express the power of agriculture in a particular locality to produce crops without regard to whether that power is due to the bounty of nature or to the efforts of man". - Den Ausgangspunkt für die Untersuchung bildete die Beobachtung, daß 168 von den 180 verschiedenen Korrelationskoeffizienten zwischen den Ernteergebnissen (10 Fruchtarten, 4 Jahre) positiv waren: "It is well known that an area producing high yields of wheat will, as a rule, produce yields of other cereals which are better than the average - in other words, the yield of wheat is positively correlated with the yields of other cereals. The starting point of this inquiry was the discovery that a similar relationship exists between county yields of the principal crops which are grown ... in England"

jedoch unabhängig von ihren Anteilswerten und Anbaumengen in den einzelnen Bezirken - aufzufinden und die 48 Bezirke dementsprechend zu klassifizieren. Zur Vermeidung von Verzerrungen durch klimatische Einflüsse wurden die Zahlen für mehrere Jahre, nämlich für 1925, 1930, 1935 und 1936 herangezogen. Werden die 10 Fruchtarten als Variable ($n=10; N=48$) aufgefaßt, so kann man die beschriebene Aufgabenstellung, für jedes Jahr getrennt, als Problem der Komponentenanalyse betrachten. Hierzu geht KENDALL davon aus, daß die erste Hauptkomponente - in Anlehnung an die geometrische Analogie als "Produktivitätsachse" bezeichnet - wegen ihrer Eigenschaften (maximale Erfassung der Variation in den Ernteergebnissen aller Fruchtarten von Bezirk zu Bezirk) als Mittel zur Bestimmung des gesuchten Maßes verwendbar ist. Als "Produktivitätskoeffizienten" P_j ($j=1,2,\dots,48$) werden, in unseren Symbolen des Abschnitts 301, die N Ausdrücke

$$(186) \quad P_j = z'_{1j} = l_{11}x_{1j} + l_{21}x_{2j} + \dots + l_{10,1}x_{10,j} \\ (j=1,2,\dots,48)$$

definiert ²⁵⁸⁾. Wenn wir uns erneut der geometrischen Veranschaulichung zuwenden, so sind die P_j die Projektionen der N Punkte auf die erste Komponentenachse, die "Produktivitätsachse", im " n_1 -Raum" ²⁵⁹⁾; die relative Größe der P_j untereinander kann daher als allgemeiner Produktivitätsindex gedeutet und eine entsprechende Einteilung der 48 Bezirke vorgenommen werden ²⁶⁰⁾. Selbstverständlich unterliegt

258) KENDALL, M.G. (1939), a.a.O., Gleichung (8) auf Seite 27. (186) entspricht bei voller Ausschreibung nach N der Gleichung (152), Seite 89. Über (155) kann der Anschluß an die Matrix $\mathcal{Z}$, Seite 16, und das System (111'), Seite 53, hergestellt werden.

259) Vgl. Seite 90-91.

260) KENDALL, M.G. (1939), a.a.O., Tabelle 5 auf Seite 28. Aus der Häufigkeitsverteilung der P_j wurden mit Hilfe des Medians und der Quartile vier gleichmäßig besetzte Größenklassen ermittelt, nach denen die Bezirke als "good", "moderate", "indifferent" und "poor" eingeteilt werden konnten.

diese Rangeinteilung einer Reihe von Vorbehalten, unter denen die Fragen der Zuverlässigkeit des untersuchten Zahlenmaterials, des Hinreichens der ausgewählten Fruchtarten und der Zweckmäßigkeit der Bezirksgliederung in Bezug auf die Zusammensetzung der Anbauflächen zu nennen sind. Auch müßte geprüft werden, ob die Erfassung der Gesamtvariation durch die erste Hauptkomponente für die gezogenen Schlüsse wirklich ausreichend ist ²⁶¹⁾.

Als Indizes der Produktivität - mit der hier verstandenen und durch die genannten Vorbehalte eingeschränkten Bedeutung - können die Koeffizienten P_j insofern angesprochen werden, als durch die Größen l_{ij} ($i=1,2,\dots,10$) ein festes Wägungsschema gegeben ist, mit dessen Hilfe nach den üblichen Formeln ein Index im eigentlichen Sinne (Wahl einer Basisperiode, unter Umständen auch Verwendung von Maßzahlen anstelle der in Mengen je Flächeneinheit ausgedrückten Werte x_{ij}) berechenbar ist. KENDALL hat drei weitere Verfahren zur Gewichtung der Ernteergebnisse entwickelt und mit der Methode des Produktivitätsindex verglichen; es handelt sich um je ein Schema für einen "Rangkoeffizienten" (ranking coefficient), einen "Geldwertkoeffizienten" (money-value coefficient) und einen "Nährwertkoeffizienten" (energy coefficient). Von Interesse ist besonders der erste Koeffizient, der als arithmetisches Mittel der Rangziffern der Fruchtarten, beginnend mit dem höchsten Ernteergebnis, in den einzelnen Bezirken definiert wird. Zwischen den P_j und den "Rangkoeffizienten" ergab sich ein so enger Zusammenhang, daß man das Verfahren der Rangziffern ganz allgemein als gute

261) Hierüber fehlen in der Originalarbeit nähere Angaben. Bei der späteren Erwähnung gibt KENDALL (1957, a.a.O., Seite 18-19) an, daß sich für 1925 ein Wert von $\lambda_1 = 4,76$ ergibt. Das entspricht nach (151'), Seite 90, für ($n=10$) einem recht geringen Varianzanteil der ersten Hauptkomponenten von 47,6%.

Approximation an die erste Hauptkomponente verwenden kann ²⁶²⁾. Für den "Geldwertkoeffizienten" und den "Nährwertkoeffizienten" wurde die Gewichtung mit den Preisen und dem Stärkegehalt der Fruchtarten vorgenommen, die Ergebnisse befanden sich mit denen der ersten beiden Koeffizienten nur teilweise in Übereinstimmung ²⁶³⁾.

Wir wollen noch anfügen, daß KENDALL seine Arbeit mit einigen Bemerkungen zum Vergleich zwischen der hier dargebotenen Form der Komponentenanalyse und der (psychologischen) Faktoranalyse abschließt. Dabei wird insbesondere vor der Versuchung gewarnt, aus der Analyse der Ernteergebnisse auf einen gemeinsamen Faktor "Produktivität" zu schließen. Die landwirtschaftliche Produktivität ist eine begriffliche Abstraktion, hinter der sich für Vereinfachungen dieser Art eine viel zu große Zahl von zusammenwirkenden Ursachen verbirgt. Man wird vielmehr auch in dieser Untersuchung, unseren Ausführungen im Abschnitt 311 entsprechend, in erster Linie das eigentliche Anliegen der Komponentenanalyse im Auge haben müssen, nämlich ²⁶⁴⁾;

262) Nach KENDALL(1939, a.a.O., Seite 43) erhält man bei Berechnung des Rangkorrelationskoeffizienten ρ zwischen den P_j und den "Rangkoeffizienten" für die vier betrachteten Jahre die Werte 0,95(1925), 0,96(1930), 0,97(1935) und 0,95(1936). Vgl. auch KENDALL(1957), a.a.O., Seite 35-36 (A ranking approximation to the first component). Es kann bewiesen werden, daß die durchschnittliche SPEARMANSCHES Rangkorrelation zwischen den Summen der Rangziffern je Fruchtart, die das n-fache der "Rangkoeffizienten" betragen, und den n einzelnen Rangordnungen ein Maximum ist. Siehe KENDALL, M.G.(1948) - Rank Correlation Methods. London: Charles Griffin & Co., Seiten 87-88 und 97-98.

263) KENDALL, M.G.(1939), a.a.O., Tabelle 14 auf Seite 39.

264) KENDALL, M.G.(1957), a.a.O., Seite 19.

"The problem here was to reduce the 10-dimensional variation to a one-dimensional variation so that the resulting variate might be used as a measure of productivity. The attempt was partly successful, and as successful as any attempt can be which endeavours to express productivity as due to the variation of one component only."

Eine von BANKS durchgeführte Bearbeitung desselben Zahlenmaterials mit gewissen Methoden der Faktoranalyse führte ²⁶⁵⁾ für 1925 zu fünf gemeinsamen Faktoren, unter denen sich wieder ein "allgemeiner Produktivitätsfaktor" mit 45,9% der Gesamtvarianz befand, während die übrigen (35,4% der Varianz) mit zahlreichen negativen Faktorladungen als "spezielle Produktivitätsfaktoren" für einzelne Fruchtarten gedeutet wurden. Auch hier wird betont, daß der Charakter dieser Faktoren lediglich im Sinne eines quantitativen Ausdrucks für Hypothesen, die auf nicht-statistischer Basis zu überprüfen sind, zu interpretieren ist ²⁶⁶⁾.

Zwei weitere Anwendungsbeispiele auf dem Gebiet der Indexzahlen stammen von TINTNER ²⁶⁷⁾. Für drei Indizes der industriellen Produktion in den USA (1935-1939=100), nämlich X_1 = Investitionsgüter, X_2 = Konsumgüter und X_3 = Bodenschätze (minerals) sowie einen Index X_4 = Produktion landwirtschaftlicher Erzeugnisse ergaben die Reihen der Periode 1919-1939 die nachstehende Korrelationsmatrix.

Tabelle 16: Korrelationskoeffizienten zwischen vier Produktionsindizes 1919-1939 nach TINTNER

	X_1	X_2	X_3	X_4
X_1	1	0,495941	0,872836	0,481240
X_2	0,495941	1	0,768279	0,709807
X_3	0,872836	0,768279	1	0,712358
X_4	0,481240	0,709807	0,712358	1

265) BANKS, C. (1954) - "The factorial analysis of crop productivity", J. Roy. Stat. Soc. B. 16, Seite 100-111.

266) Vgl. Seite 119.

267) TINTNER, G. (1946), a.a.O., Seite 482-485.
 - - - Ders. (1952), a.a.O., Seite 110-114.

Der größte Eigenwert dieser Matrix ist mit $\lambda_1 = 3,033424$ bestimmt worden, woraus sich ein Varianzanteil der ersten Hauptkomponenten von $(\lambda_1/n)100 = 75,83\%$ ergibt. Die Koordinaten a_{i1} ($i=1,2,3,4$) des zugehörigen Eigenvektors a_1 sind zusammen mit ihren Quadraten, die nach (38) als Anteile der ersten Hauptkomponenten an den Varianzen der einzelnen x_i aufzufassen sind, in der folgenden Übersicht enthalten ²⁶⁸⁾.

Tabelle 17: Anteilswerte der ersten Hauptkomponenten an den Varianzen von vier Indexreihen nach TINTNER

Indexreihen (standardisiert)	Koordinaten des Vektors a_1	Varianzanteile der ersten Hauptkomponenten (in v.H.)
(1)	(2)	(3)
Investitionsgüter x_1	0,817396	66,81
Konsumgüter x_2	0,888102	78,87
Bodenschätze x_3	0,951934	90,61
Landwirtschaftliche Erzeugnisse x_4	0,818776	67,04

Etwas günstigere Ergebnisse erzielte TINTNER bei der Bestimmung der ersten Hauptkomponenten einer Gruppe von drei Indexreihen für Großhandelspreise in den USA (1926=100), ebenfalls für den Zeitraum 1919-1939. Dabei wurden die Großhandelspreise für landwirtschaftliche Erzeugnisse (wholesale farm prices) (X_5), Nahrungsmittel (X_6) und alle übrigen Güter (X_7) untersucht; aus $\lambda_1 = 2,871360$ ließ sich ein Anteil an der Gesamtvarianz von 95,71% errechnen, während die der Tabelle 17 entsprechende Aufstellung folgende Zahlen aufweist:

²⁶⁸⁾ Vgl. Tabelle 5, Seite 81.

Tabelle 18: Anteilswerte der ersten Hauptkomponenten an den Varianzen von drei Preisindizes 1919-1939 nach TINTNER

Indexreihen (standardisiert)	Koordinaten des Vektors α_1	Varianzanteile der ersten Haupt- komponenten (in v.H.)
(1)	(2)	(3)
Landwirtschaft- liche Erzeug- nisse x_5	0,987867	97,59
Nahrungsmittel x_6	0,990160	98,05
Übrige Güter x_7	0,957621	91,70

An dieses Beispiel hat TINTNER einen interessanten Vergleich mit dem Gesamtindex der Großhandelspreise (all commodity wholesale price index) des U.S. Bureau of Labor Statistics angeschlossen. Wir haben im Abschnitt 302 den von GIRSHICK behandelten Fall erörtert, daß es eine Eigenschaft der Hauptkomponenten ist, den auf die zufälligen Fehler der normierten Ausgangsvariablen x_i entfallenden Anteil an ihren Varianzen als Minimum zu enthalten und gleichzeitig die Quadratsumme R_k der Korrelationskoeffizienten mit allen Variablen x_i zu maximieren ²⁶⁹⁾. Für die drei Großhandelspreisindizes erhalten wir in Bezug auf die erste Hauptkomponente die diesen Bedingungen genügende Gleichung

$$(187) \quad z_1 = \sum_{i=5}^7 a'_{1i} x_i = 0,343693x_5 + 0,344845x_6 + 0,333508x_7.$$

Wird (187) ²⁷⁰⁾ für alle 21 Jahre der Periode berechnet

269) Vgl. Seite 92-96.

270) TINTNER, G. (1952), a.a.O., Gleichung (21) auf Seite 113. (187) entspricht (160) auf Seite 93 und im Vergleich zum Zahlenbeispiel des Abschnitts 21 der ersten Zeile von (144'), Seite 80. Die Koeffizienten a'_{ki} sind durch (167), Seite 95, mit den a_{ik} der Tabelle 18 (Spalte 2) verbunden.

und den Werten des Gesamtindex gegenübergestellt, so erhält man eine sehr hohe Korrelation von 0,991. Dieses Resultat legt die Vermutung nahe, daß die Komponentenanalyse als Instrument für die Konstruktion und Kontrolle von Gesamtindizes eine gewisse Bedeutung haben kann. Darauf kommen wir in den nächsten beiden Abschnitten zurück; die hier zugrunde gelegte Ausgangsgesamtheit ist zweifellos zu klein, um derartige Schlüsse zu rechtfertigen.

TINTNER versteht die skizzierten Beispiele mit einer "tentative economic interpretation" dahingehend, daß aus den Ergebnissen das Wirken eines "allgemeinen Produktions-" beziehungsweise "Preisbildungsfaktors" entnommen werden könne, im Gegensatz etwa zu der Möglichkeit, daß bei einer Volkswirtschaft mit geringen Beziehungen zwischen dem industriellen und dem landwirtschaftlichen Sektor im erstgenannten Beispiel die erste Hauptkomponente nur für die Variablen X_1 , X_2 und X_3 bedeutsam gewesen wäre und sich für X_4 ein "landwirtschaftlicher Produktionsfaktor" konstituiert hätte. Diese Art von faktoranalytischen Überlegungen erscheint jedoch bedenklich; wir haben auf die Gründe schon an mehreren Stellen unserer Arbeit hingewiesen und schließen uns in diesem Falle mit Nachdruck der Feststellung an ²⁷¹⁾:

"A more detailed analysis should of course be carried out not only to confirm this result with respect to the broad categories of production used here but also to apply it to the production of individual commodities."

413 Die Komponentenanalyse als Hilfsmittel zur Indexkonstruktion

Die auf den vorhergehenden Seiten geschilderten Untersuchungen von RHODES, KENDALL und TINTNER haben die Anwendung der Komponentenanalyse in einigen Bereichen der

271) TINTNER, G. (1952), a.a.O., Seite 112.

Indexzahlen zum Gegenstand. Da aber nur Einzelfragen angeschnitten wurden und der Umfang der Anwendungsmöglichkeiten nach unserer Ansicht größer ist als es bisher sichtbar wurde, wollen wir uns in diesem und dem nächsten Abschnitt einer mehr systematisch gehaltenen Zusammenstellung zuwenden.

Es ist bereits zum Ausdruck gekommen, daß man gelegentlich versucht hat, die Modellvorstellung der Faktoranalyse bei indextheoretischen Untersuchungen heranzuziehen. Dies ist bei RHODES und mit gewissen Einschränkungen auch bei TINTER der Fall. In der Tat erscheint das Problem der Indexzahlen in der Definition von BOWLEY ²⁷²⁾,

"Index-numbers are used to measure the change in some quantity which we cannot observe directly, which we know to have a definite influence on many other quantities which we can so observe, tending to increase all, or diminish all, while this influence is concealed by the action of many causes affecting the separate quantities in various ways",

im Vergleich zur Aufgabenstellung der Faktoranalyse geradezu als Musterfall einer nicht-psychologischen Anwendung. Diese könnte aber nur dann sinnvoll sein, wenn nach älteren Vorstellungen die Indexzahlen als Methode zur Bestimmung globaler Phänomene - etwa der Kaufkraft des Geldes und ihrer Veränderungen im Falle von Preisindizes - betrachtet werden ²⁷³⁾. Da nach heute überwiegender Auffassung in solchen Fragen eine Überforderung der Indextheorie zu erblicken ist und man unter Indexzahlen im wesentlichen - vielleicht mit Ausnahme gewisser Sonderprobleme der Ökonometrie - den weniger anspruchsvollen und in der Aussagekraft beschränkten, gleichwohl durch vereinfachende Modellbildung eindeutigen Spezialfall kombinierter Meßzahlen

272) BOWLEY, A.L. (1937⁶) - Elements of Statistics (Studies in Economics and Political Science No. 8). London (New York): Staples Press, Seite 196.

273) Vgl. zum Beispiel die Auseinandersetzung mit dem "Sinn der Indexziffern" als Ausdrucksmittel der Quantitätstheorie bei MEERWARTH, R. (1920) - Einleitung in die Wirtschaftsstatistik. Jena: Gustav Fischer, Seite 244-250. Siehe auch ANDERSON, O. (1957). a.a.O.. Seite 38-39.

versteht ²⁷⁴⁾, soll auch in der Anwendung der Komponentenanalyse auf umfassende faktorielle Erwägungen verzichtet werden. Dagegen wollen wir zeigen, daß die Methode der Hauptkomponenten als Hilfsmittel zur Konstruktion und Überprüfung zahlreicher Indextypen von Bedeutung sein kann.

Neben ihrer Verwendung bei echten Indexzahlen tritt die Frage auf, ob die Komponentenanalyse ein brauchbares Instrument für die Aufstellung von Größen ist, die zwar der Messung von relativen zeitlichen oder regionalen Veränderungen eines Sachverhalts dienen, aber im eigentlichen Sinne entweder keine Indizes oder nur Vorstufen zu solchen sind. Der "Produktivitätsindex" von KENDALL ist ein Beispiel hierfür, sein Indexcharakter im Sinne der Aussage ²⁷⁵⁾ unbestreitbar:

"It often happens ... that a set of similar events moving through time or space gives rise to some general concept expressing variation in their common element."

Diese Fragen werden wir im Abschnitt 414 behandeln, wo außerdem einige Betrachtungen über den Nutzen der Komponentenanalyse bei der Verwendung von Indizes in makroökonomischen Modellen anzustellen sind. Auch der Abschnitt 415, für den eine Prüfung der Einsatzmöglichkeiten der Komponentenanalyse bei gewissen Problemen der Wirtschaftsrechnungen vorgesehen ist, steht wegen der Berührungspunkte mit dem Preisindex der Lebenshaltung in einem Zusammenhang mit allgemeinen Indexproblemen. Schließlich sei wiederholt, daß die Anwendung der Komponentenanalyse in allen angedeuteten Fällen unter der doppelten Zielsetzung der Mithilfe bei der Bildung von Indexzahlen sowie der Überprüfung der

274) PFANZAGL, J. (1960) - Allgemeine Methodenlehre der Statistik I. Elementare Methoden unter besonderer Berücksichtigung der Anwendungen in den Wirtschafts- und Sozialwissenschaften (Sammlung Götschen, Bd. 746/746a). Berlin: Walter de Gruyter, Seite 121-123.

275) YULE-KENDALL (1950), a.a.O., Seite 590. Die Untersuchung über den Produktivitätsindex haben wir auf Seite 141-145 behandelt, auf ähnliche Größen kommen wir im Abschnitt 421 zurück.

adäquaten Wiedergabe der jeweils betrachteten ökonomischen Sachverhalte durch bereits berechnete Indizes erfolgen kann.

Ein ganz generelles Problem der Konstruktion von Indexzahlen wird durch die Forderung hervorgerufen, daß ein komplexer Vorgang der wirtschaftlichen Realität durch eine einzige Reihe von zeitlich oder regional variablen Größen dargestellt werden soll. Da aber aus Gründen der mangelnden Eindeutigkeit derartig umfassender Fragestellungen sowie der im allgemeinen beschränkten Erhebungsmöglichkeiten nur ein Teil der zu indizierenden Tatbestände berücksichtigt werden kann, muß der Indexberechnung die Bildung eines vereinfachten Modells vorausgehen; dieses soll den Index einerseits möglichst nahe an den "idealen Index" heranführen, andererseits die genügende Überschaubarkeit seiner praktischen Handhabung gewährleisten. Der Schlüssel zu dieser Modellbildung liegt in der repräsentativen Auswahl und Gewichtung einzelner Größen aus der Gesamtheit aller in den Bereich des jeweiligen Index fallenden ökonomischen Tatbestände ²⁷⁶⁾. Nehmen wir das Beispiel des in der Bundesrepublik Deutschland monatlich berechneten Index der Einzelhandelspreise, so sind neben der Wahl der Indexformel (LASPEYRES, Mittelwert aus Meßzahlen), der Basisperiode (Jahresdurchschnitt 1950=100 %) und des Wägungsschemas (Umsatzwerte des Einzelhandels 1950) die Festsetzung des Preisschnitts und die Bestimmung der einzubeziehenden Waren ²⁷⁷⁾ methodisch

-
- 276) Zur Modellbildung und Standardisierung vgl. KELLERER, H. (1960) - Statistik im modernen Wirtschafts- und Sozialleben (rowohlts deutsche enzyklopädie, Bd. 103/104). Hamburg: Rowohlt, Seite 76-78. Bezogen auf Preisindizes sind ferner in diesem Zusammenhang die ausführlichen Erörterungen von MEERWARTH, R. (1925) - Nationalökonomie und Statistik (Handb. d. Wirtsch.-u. Sozialwiss., Bd. 7). Berlin (Leipzig): Walter de Gruyter & Co., Seite 396-444, zu nennen.
- 277) DENEFFE, P.; KELLER (1952) - "Zur Berechnung eines Index der Einzelhandelspreise", Wirtsch. u. Stat. N.F. 4, Seite 363-369. Zur allgemeinen Orientierung vgl. auch UNITED NATIONS, ECONOMIC AND SOCIAL COUNCIL, STATISTICAL COMMISSION (Hrsg.) (1958) - General-Purpose Index Numbers of Retail Prices (Doc. E/CN.3/247).

entscheidend. Für das letztgenannte Problem wäre eine Anwendung der Komponentenanalyse unter folgenden Gesichtspunkten denkbar. Das Gewicht jeder einzelnen Branche ist als anteiliger Umsatz am Gesamtumsatz des Basisjahres definiert, jedoch muß darüber hinaus innerhalb der Branche selbst eine Auswahl und Gewichtung der für diese repräsentativen Warenpositionen erfolgen. So erhält in der Hauptbranche "Hausrat und Wohnbedarf" die Branche "Geschäfte für Möbel" die Gewichtszahl 301 v.T. am Index der Hauptbranche und 31 v.T. am Gesamtindex; die Anzahl der Waren beträgt 11²⁷⁸⁾. Die Frage ist nun, inwieweit diese Warenpositionen als repräsentativ für das gesamte Sortiment des Möbelhandels anzusehen sind und wie das festgelegte Gewicht von 31 v.T. auf sie zu verteilen ist²⁷⁹⁾. Hierzu könnte man zur objektiven Ermittlung der bestmöglichen Repräsentation die in N Perioden oder Regionen²⁸⁰⁾ festgestellten Umsätze etwa sämtlicher oder jedenfalls der meisten für den Einzelhandel mit Möbeln typischen Produkte als n Variable

-
- 278) DENEFFE-KELLER(1952), a.a.O., Tabelle 1 auf Seite 364. Übrigens kann auch das Wägungsschema als solches der Überprüfung durch eine Komponentenanalyse unterzogen werden. Vgl. dazu die später folgenden Ausführungen zur Kombination von Gruppenindizes auf Seite 158.
- 279) Dabei ist zu beachten, daß wegen des weitaus größeren Sortiments die Umsatzanteile der 11 ausgewählten Waren jedenfalls nicht 31 v.T. des Gesamtumsatzes erreichen, vielmehr muß eine Zuordnung der Umsatzanteile der weniger bedeutenden Positionen erfolgen. Dieselbe Problematik ergibt sich bei der Gliederung nach Art und Herkunft der Waren, wo zum Beispiel die durch den Einzelhandel umgesetzten Getränke bei einem Gewicht von 24 v.T. durch 7 Warenpositionen repräsentiert werden. Vgl. auch PFANZAGL, J. (1960), a.a.O., Seite 128-129.
- 280) Bei zeitlicher Betrachtungsweise wäre infolge der vermutlich begrenzten Anzahl der zur Verfügung stehenden Perioden in aller Regel der ungewöhnliche, aber methodisch durchführbare Fall ($n > N$) gegeben (vgl. Seite 12-13). Strukturell aufschlußreicher erscheint jedoch die regionale Aufteilung im Zeitpunkt der Basisperiode, wobei einzelne Geschäfte, Durchschnittsergebnisse von Städten oder noch größeren Gebietseinheiten als Gliederungsmerkmale gewählt werden können.

auffassen und diese nach entsprechender Normierung einer Komponentenanalyse unterziehen. Für die Verwendung der Hauptkomponenten, von denen erfahrungsgemäß eine kleine Zahl zur fast vollständigen Erfassung der Gesamtvariation der Warenumsätze genügt, bieten sich zwei Wege an. Ihre Berechnung nach (111) ermöglicht zwar die Wiedergabe der zeitlichen oder regionalen Variation in den Umsätzen mit p oder nur einigen der ersten unter den p "neuen" Variablen und wird insofern der Zielsetzung gerecht; jedoch sind die Hauptkomponenten abstrakte Zahlen und keine Umsatzreihen realer Güter, für die die Preise feststellbar sind. Dagegen wird es im Einzelfall möglich sein, der Größenordnung der Koeffizienten a_{ik}/λ_k - analog zur Faktorstruktur der Faktoranalyse - besonders enge und durch Korrelationsrechnung nachprüfbare Zusammenhänge zwischen je einer im Varianzbeitrag bedeutsamen Hauptkomponenten z_k und einer der Ausgangsvariablen x_i zu entnehmen, so daß letztere als für die Umsatzstruktur repräsentative Waren in die Auswahl genommen werden können. Der zweite Weg führt über das System (11), der Umkehrung²⁸¹⁾ von (111), im Grunde zu demselben Ergebnis: Es werden diejenigen x_i nacheinander als repräsentativ betrachtet, die die höchsten Werte a_{ik} in Bezug auf die für die Gesamtvariation ausreichenden $z_1, z_2, \dots$ aufweisen. Als Gewichte der ausgewählten Variablen wird man ihre Anteile an der Gesamtvariation nach (105) verwenden, die entsprechenden v.H.-Werte²⁸²⁾ lassen sich leicht in die durch das Wägungsschema gegebenen Anteile am Hauptbranchen- oder Gesamtumsatz umrechnen. Mag der vorgeschlagene Weg auch rechentechnisch aufwendig und hinsichtlich

281) Vgl. Seiten 6 und 53. Da es sich bei der Gewichtung des Index der Einzelhandelspreise um Umsatzwerte, nicht Umsatzmengen handelt, kann die Normierung der Variablen auf die Bildung der Differenzen vom Mittelwert beschränkt werden. Allerdings wären damit einige Modifizierungen unserer Methode erforderlich.

282) Vgl. Seite 51-52.

des erforderlichen Zahlenmaterials recht anspruchsvoll sein, so stellt er doch eine exakte Methode zur Bestimmung der Umsatzstruktur dar, deren Überprüfung im Zusammenhang mit den derzeitigen Verfahren ²⁸³⁾ empfehlenswert erscheint. Unter der Voraussetzung relativ stabiler Verbrauchsgewohnheiten der Einzelhandelskunden genügt es, das Verfahren nur jeweils zum Zeitpunkt eines Basiswechsels für den Index anzuwenden; als besonderer Vorteil ist die Möglichkeit hervorzuheben, daß die aufeinanderfolgende Maximierung der Varianzanteile zu einer Verringerung der Warenpositionen - für den Index der Einzelhandelspreise werden zur Zeit 655 Waren berücksichtigt - und damit zu einer kleineren Zahl von Preisreihen führen kann. Auch für die laufende Beobachtung der Preise ²⁸⁴⁾, bei der die Ermittlung eines für den gesamten Geltungsbereich des Index verbindlichen Durchschnitts

283) DENEFFE-KELLER(1952), a.a.O., Seite 304: "Schwieriger gestaltete sich die Auswahl und Gewichtung der einzelnen Waren innerhalb der Branchen. Hierzu wurden frühere amtliche Untersuchungen des Reichskuratoriums für Wirtschaftlichkeit über die Betriebsstruktur im Einzelhandel, ferner Erhebungen und Mitteilungen von Handelsverbänden und einzelnen Firmen, Ergebnisse der Produktionsstatistiken und der Verbrauchsstatistiken, insbesondere der Wirtschaftsrechnungen in Arbeitnehmerhaushalten, herangezogen."

284) Vgl. die für Großhandelspreise beschriebenen Verfahren in UNITED NATIONS, ECONOMIC AND SOCIAL COUNCIL, STATISTICAL COMMISSION(Hrsg.)(1960) - Problems and Methods in the Gathering of Representative and Comparable Wholesale Price Series (Doc. E/CN.3/264). Ein gutes Beispiel für die Anwendung auf Untersuchungen von Preisstrukturen ist eine von WAGENFÜHR erwähnte Erhebung von 20 000 Preisen für 200 Artikel, die 1954 für den Vergleich des Lebenshaltungsniveaus in den Montanunionländern durchgeführt wurde:"Zur systematischen Durchforschung der Preisstruktur haben wir versucht, eine Reihe von Methoden versuchsweise zu entwickeln. Dabei handelte es sich etwa um folgende Fragestellungen: Erstens die Frage, kann man mit Leitwaren arbeiten, d.h. gibt es Schlüsselerzeugnisse, die sozusagen die Preisverhältnisse in den einzelnen Ländern widerspiegeln?". Dieses Problem würde sich vermutlich mit Hilfe der Hauptkomponenten der Preisvariationen zweckmäßig lösen lassen. Vgl. WAGENFÜHR, R.(1958) - "Probleme des internationalen Vergleichs des Lebensniveaus", Allg.Stat.Arch. 42, Seite 307-308.

aus zahlreichen Einzelpreisen methodisch bedeutend ist, wäre der Einsatz der Komponentenanalyse als Hilfsmittel zur Untersuchung der zeitlichen und vor allem regionalen Variationen des Preisgefüges denkbar.

Die auf den Index der Einzelhandelspreise bezogene Prüfung einiger Anwendungsmöglichkeiten der Komponentenanalyse kann auf eine ganze Reihe von anderen Indizes ausgedehnt werden, wobei spezielle, jedoch nicht grundsätzliche Unterschiede auftreten. Als weitere Beispiele nennen wir den Index der Erzeugerpreise landwirtschaftlicher Produkte ²⁸⁵⁾ mit der besonderen Problematik wechselnder Gewichtung in den Saisonphasen sowie den Erzeugerpreisindex industrieller Produkte, dessen 1078 Einzelpositionen ganz im Sinne der bereits skizzierten Verwendung der Komponentenanalyse unter den Gesichtspunkten bestimmt wurden ²⁸⁶⁾,

"... daß (einmal) mit den ausgesuchten Waren repräsentativ die Preisentwicklung der nicht im Index erfaßten Erzeugnisse mit dargestellt wird, zum anderen ... die für eine Industriegruppe oder einen Industriezweig ausgewählten Waren eine möglichst große Bedeutung am Umsatz der entsprechenden Industriegruppe bzw. des entsprechenden Industriezweigs haben und so im Index die durchschnittliche Preisentwicklung der gesamten Gruppe oder des gesamten Zweigs wiederzugeben im Stande (sind)."

Wir haben bisher nur Preisindizes genannt und müssen ergänzend hinzufügen, daß die Komponentenanalyse auch für strukturelle Untersuchungen im Zusammenhang mit der Konstruktion von Mengenindizes in Betracht kommt. Hier ist

285) DENEFFÉ, P.; BRANDNER, H. (1952) - "Erzeugerpreise landwirtschaftlicher Produkte", Wirtsch. u. Stat. N.F. 4, Seite 61-65. In Übereinstimmung damit der Index forstwirtschaftlicher Produkte, vgl. BRANDNER, H. (1958) - "Neuberechnung eines Index der Erzeugerpreise forstwirtschaftlicher Produkte", Wirtsch. u. Stat. N.F. 10, Seite 498-501.

286) BRANDNER, H. (1953) - "Der Erzeugerpreisindex industrieller Produkte. Endgültige Berechnung", Wirtsch. u. Stat. N.F. 5, Seite 251. Nicht erwähnt wurde der Preisindex der Lebenshaltung, für den ebenfalls Anwendungsmöglichkeiten (z.B. repräsentative Füllungen des "Warenkorbes") gegeben sind. Vgl. Abschnitt 415 und FÜRST, G.; DENEFFÉ, P. (1952) - "Der neue Preisindex der Lebenshaltung", Wirtsch. u. Stat. N.F. 4, Seite 439-455.

in erster Linie der Index der industriellen Nettoproduktion hervorzuheben, bei dem die einbezogenen Produktionsausstoßmengen "... möglichst gut über den betrachteten Bereich (Industriezweig oder gesamte Industrie) gestreut" sein sollen und die für sie als Gewichte verwendeten Nettoproduktionswerte "... annäherungsweise aus den Nettoproduktionswerten der Industriezweige gewonnen" werden ²⁸⁷⁾. Auch andere Mengenindizes, etwa solche der landwirtschaftlichen Produktion oder des Verbrauchs von Nahrungsmitteln, weisen methodische Probleme auf ²⁸⁸⁾, die durch den Einsatz der Komponentenanalyse einer verbesserten Lösung zugeführt werden können.

414 Kombinationen von Indexzahlen

In der Praxis der Aufstellung von Indizes ist es vielfach üblich, Indexzahlen für größere wirtschaftliche Bereiche in Gruppen- oder Teilindizes zu zerlegen oder, in umgekehrter Richtung, eine Reihe von einzelnen Indizes zu General- oder Gesamtindizes zu vereinigen. So werden bei dem bereits erwähnten Index der Einzelhandelspreise neben den Gesamtergebnissen weitere 55 Gruppenindizes, davon 37 für Nahrungsmittel, Bekleidungsgegenstände und einige andere Waren nach Art und Herkunft, die übrigen für die Entwicklung der Einzelhandelspreise nach ausgewählten Branchen, veröffentlicht, während der Index der industriellen Nettoproduktion in 65 über- und nebengeordnete Branchenindizes gegliedert

-
- 287) WERNER, K. (1958) - Die Industriestatistik der Bundesrepublik Deutschland. Berlin: Duncker & Humblot, Seite 126. Vgl. auch RÜDIGER, W. (1956) - "Neuberechnung des Index der industriellen Nettoproduktion", Wirtsch. u. Stat. N.F. 8, Seite 129-146.
KELLERER, H. (1960), a.a.O., Seite 90-91.
PFANZAGL, J. (1960), a.a.O., Seite 144-145.

- 288) Siehe zum Beispiel ORGANISATION FOR EUROPEAN ECONOMIC CO-OPERATION (Hrsg.) (1955) - The Measurement of Agricultural Production and Food Consumption.

ist ²⁸⁹⁾. Diese Aufteilung beziehungsweise Kombination von Indexzahlen stellt formal nur eine einfache Erweiterung der gewöhnlichen Indexkonstruktion durch gewogene Mittelwertbildung aus den Gruppenindizes, wobei die Gewichte dem jeweiligen Gesamtwägungsschema entnommen werden, dar ²⁹⁰⁾, jedoch eröffnet sich hier eine zusätzliche Anwendungsmöglichkeit für die Komponentenanalyse. Man kann nämlich Indizes auch dadurch zu einem Gesamtindex vereinigen, daß n einzelne Indexreihen - im wesentlichen nach ihrem zeitlichen Verlauf durch N Perioden, aber auch in regionaler Gliederung - als Variable x_i ($i=1,2,\dots,n$) aufgefaßt und mit der Methode der Hauptkomponenten analysiert werden. Die gewünschte Globalgröße ergibt sich durch ²⁹¹⁾

$$(188) \quad z_1 = \frac{1}{\lambda_1} \sum_{i=1}^n a_{i1} x_i$$

als erste Hauptkomponente mit einem maximalen Anteil an der Gesamtvariation in den Gruppenindizes. Das Resultat einer solchen Untersuchung läßt sich nicht als Gesamtindex im eigentlichen Sinne bezeichnen, weil die Summe der

-
- 289) STATISTISCHES BUNDESAMT(Hrsg.) - Statistischer Bericht, Arb.-Nr. VI/3/123 vom 29.März 1960, Seite 6-7.
- - - - -Dass. (Hrsg.) - Die Industrie der Bundesrepublik Deutschland. Reihe 2: Produktion ausgewählter Erzeugnisse, Produktionsindex, Februar 1960. Wiesbaden: W.Kohlhammer, Seite 22-27.
- 290) Vgl. den Abschnitt über Sub-Indizes bei PFANZAGL,J.(1960), a.a.O., Seite 126-127.
- 291) Vgl. (111'') auf Seite 53. Zu beachten ist, daß die x_i und z_k wiederum als normierte Variable zu betrachten sind. Soweit z_1 die Varianzen der Variablen nur ungenügend erfaßt, beeinträchtigt dies die Aussagefähigkeit des gesuchten "Index". Eine Kombination der dann für eine vollständige Erfassung notwendigen $z_1, z_2, z_3 \dots$ zu einer einzigen Zahl ist definitionsgemäß nicht möglich, ihre Zusammenfassung durch Mittelwertbildung oder auf anderem Wege methodisch fragwürdig.

durch (188) gegebenen Gewichte, nämlich

$$(189) \quad \frac{1}{\lambda_1} (a_{11} + a_{21} + \dots + a_{n1}) = \frac{\sum_{i=1}^n a_{i1}}{\sum_{i=1}^n a_{i1}^2},$$

nicht gleich 1 ist. Man kann aber sagen, daß der sachliche Zweck einer Kombination von Indexzahlen auf diesem Wege gut erfüllt wird. Im übrigen besteht die Möglichkeit, die normierten "Indizes" z_{1j} ($j=1,2,\dots,N$) ihrerseits unter Abstimmung mit der Basisperiode der Teilindizes in Meßzahlen umzurechnen. Weiterhin kann durch vollständige Berechnung der Gleichungssysteme (11) und (111) ein Einblick in die Struktur der Zusammenhänge zwischen dem Gesamtindex und jedem einzelnen der Teilindizes gewonnen werden.

Für die praktische Anwendung von Indexzahlen vom Typ (188), die in ihrem Aufbau dem "Produktivitätsindex" von KENDALL entsprechen und denen wir im Abschnitt 421 erneut begegnen werden, kommt im Bereich der amtlichen Statistik neben der vermutlich weniger interessanten Zusammenfassung von Teilindizes die Kontrolle bereits berechneter Gesamtindizes in Frage, wobei nicht nur Gruppenindizes sondern auch größere Gesamtheiten von Meßzahlreihen einzelner Preise oder Mengen als Berechnungsgrundlagen dienen können. Diesen Gesichtspunkt hat TINTNER bei den im Abschnitt 412 diskutierten Anwendungsbeispielen erwähnt²⁹²). Eine direkte Verwendungsmöglichkeit besteht bei der Konstruktion von Indexzahlen, die auf übernationaler Ebene bestimmt werden und sich auf die Ergebnisse nationaler Indexberechnungen stützen. Dabei wird es oft vorkommen, daß die durch unterschiedliche Klassifizierungen und Berechnungsmethoden in den einzelnen

292) TINTNER, G. (1952), a.a.O., Seite 110: "How perfect is the representation of all the various prices, for instance, by some general price index? Which percentage of the variance of various quantities produced in the economy is accounted for by a certain production index?"

Ländern bedingten Abweichungen zwischen den Teilindizes eine Zusammenfassung auf gewöhnlichem Wege erschweren. Hier liegt der Nutzen einer Komponentenanalyse in der Berechnung einer angenäherten Indexzahl, die ausschließlich die Variationen in den Teilindizes berücksichtigt und von den verfahrensmäßigen Differenzen unabhängig ist ²⁹³⁾.

Wir haben schon darauf hingewiesen, daß die Zusammenfassung von Gruppenindizes durch die Komponentenanalyse auch auf die Kombination von variablen Preis- oder Mengenreihen ausgedehnt werden kann, die nicht selbst schon Indexzahlen sind. Ein solches Vorgehen führt wegen des Wägungsschemas (189) zu einer im Grunde völlig neuen Methode der Indexkonstruktion, die nach Auffassung einiger Autoren für die statistische Bewältigung des Aggregationsproblems ²⁹⁴⁾ in der Ökonometrie von Interesse ist. Hierzu schreibt TINTNER ²⁹⁵⁾:

"Many questions of economic policy require a knowledge of the broad economic relationships which are discussed in economic theory under the name of general equilibrium. This is true, for instance, of problems of full employment, taxation, subsidies, etc., which ought to be discussed in the most general terms possible. It is obviously impossible to verify statistically a complete general

293) Als Anregung zu entsprechenden Untersuchungen können zum Beispiel die folgenden Veröffentlichungen dienen:

UNITED NATIONS, STATISTICAL OFFICE(Hrsg.)(1958) - Methods Used in Compiling the United Nations Price Indexes for Basic Commodities in International Trade (Stat.Pap., Ser.M., No. 29).

KLAYMAN,M.J.(1960) - "International index numbers of food and agricultural production", Monthly Bull. Agr.Econ.Stat. 2, Seite 12-14.

294) Vgl. etwa für aggregierte Preise und Mengen WOLD,H.; JUREEN,L.(1953) - Demand Analysis. New York: J.Wiley & Sons (London: Chapman & Hall), Seite 108-110. Siehe ferner KENDALL-BUCKLAND(1957), a.a.O., Seite 6-7.

295) TINTNER,G.(1952), a.a.O., Seite 109-110. Vgl. auch - - -Ders.(1956) - "Statistik und Ökonometrie", Stat. Vierteljahresschr. 2, Seite 93-94, wo die "Methode der prinzipiellen Komponente" (!) von HOTELLING zur Behandlung des Aggregationsproblems vorgeschlagen wird. Siehe außerdem STONE,R.(1947), a.a.O., Seite 32.

equilibrium system because of the great number of variables involved. It would be necessary to include all prices of all commodities produced and consumed, all interest rates, etc. It is obvious that such a procedure would involve literally thousands and possibly millions of variables. Hence it seems to be necessary to substitute certain 'indices' for groups of these variables."

Eine tatsächliche Verwendung der Komponentenanalyse unter diesen Gesichtspunkten liegt nach unserem Wissen noch nicht vor, jedoch erscheint es nicht ausgeschlossen, daß im Hinblick auf die zahlreichen methodischen Probleme des zweck- und systemgerechten Einsatzes von Indexzahlen in makroökonomischen Kreislaufmodellen und volkswirtschaftlichen Gesamtrechnungen ²⁹⁶⁾ "Indizes" des Typs (188) - das Vorhandensein entsprechender Großrechenanlagen vorausgesetzt - zumindestens in erster Annäherung ein gutes Hilfsmittel zur numerischen Darstellung gesamtwirtschaftlicher Prozesse sind.

415 Anwendungen auf Wirtschaftsrechnungen: Untersuchung der Verbrauchsstruktur

Den Abschluß unserer Ausführungen über Indexzahlen soll eine Betrachtung darüber bilden, wie die Komponentenanalyse bei einigen Problemen der Verbrauchsforschung verwendet werden kann. Diese Fragen sind auch wegen der Bedeutung der Wirtschaftsrechnungen für den Preisindex der Lebenshaltung von Interesse. Es liegen hierzu zwei Arbeiten vor, auf die wir nachstehend eingehen wollen.

296) BARTELS, H.; FÜRST, G. (1949) - "Preisindices im volkswirtschaftlichen Güterkreislauf", Wirtsch. u. Stat. N.F. 1, Seite 261-267.
STONE, R. (1956) - Quantity and Price Indexes in National Accounts. Paris: Organisation for European Economic Co-Operation, insbesondere Seite 97-106.
UNITED NATIONS, ECONOMIC AND SOCIAL COUNCIL, STATISTICAL COMMISSION (Hrsg.) (1957) - A System of Price and Quantity Indexes for National Accounts (Doc. E/CN.3/L.46), insbesondere Kapitel 4 und 5, Seite 32-81.

POTI 297) analysierte die Struktur der Verbrauchsausgaben von 497 Mittelstandsfamilien, die in Kalkutta auf Stichprobenbasis ausgewählt wurden. Die Untersuchung fußte auf der Feststellung, daß die Einteilung der Familien nach der Höhe der Verbrauchsausgaben die Variationen in den vier Ausgabengruppen lebensnotwendige Nahrungsmittel, andere Nahrungsmittel, Mieten und Heizmaterial sowie Kleidung und sonstige Ausgaben nicht genügend zu erklären schien. Eine Komponentenanalyse mit ($n=3; N=497$) 298) ergab drei Hauptkomponenten, deren Varianzanteile 84,4%, 15,5% und 0,1% betragen. Dieses Resultat ermöglichte neben der üblichen Funktion des Ersetzens der drei Ausgangsvariablen durch eine Hauptkomponente - eines für die Klassifizierung der Haushalte im Sinne eines "Armutindex" (poverty level) nützlichen Ergebnisses, da z_1 mit der Familiengröße zu- und mit der Gesamtausgabe abnahm - und einer nach unserer Auffassung nicht unbedenklichen quasi-faktoriellen Interpretation von z_1 299) den zusätzlichen Schluß, daß für eine Unterscheidung nach der Ausgabenstruktur die Höhe der Gesamtausgaben mit einem Varianzanteil von 10,82% sowie die Familiengröße, deren Korrelation mit z_1 nur -0,329 beträgt, ungeeignet sind. Deshalb wird vorgeschlagen, als Kriterium für Variationen in der Ausgabenstruktur der Familien die erste Hauptkomponente, von POTI "most discriminating function" genannt, zu verwenden. Eine etwas eingehendere Untersuchung der Verbrauchsausgaben von 415 Familien in acht Städten der Vereinigten Staaten führte

297) POTI, S.J. (1946) - "A study of expenditure patterns of Calcutta middle class family budgets", Sankhyā 7, Seite 425-428.

298) Da die Summe der vier angegebenen Ausgabengruppen den Gesamtausgaben entspricht, konnte die Untersuchung auf beliebige drei der Gruppen beschränkt werden. Allerdings wird nicht angegeben, welche Gruppen in die Auswahl kamen.

299) POTI, S.J. (1946), a.a.O., Seite 428: "Since the principal component ... explains a major portion of the variance, there exists a single economic factor which influences to a great extent the expenditure pattern."

MILLICAN durch ³⁰⁰⁾. Hierbei bildeten 20 Bedarfsgruppen (Nahrungsmittel, Bekleidung, Wohnung und weitere) und fünf Merkmale der Familienstruktur ³⁰¹⁾ (Familieneinkommen, Familiengröße, Alter und Berufsausbildung des Haushaltsvorstands sowie geographischer Standort) eine Ausgangsgesamtheit von 25 Variablen ($n=25; N=415$), auf die die Schwerpunktmethode der Faktoranalyse mit anschließender Rotation durch das Quartimax-Verfahren ³⁰²⁾ angewendet wurde. Man erhielt fünf gemeinsame Faktoren, deren Deutung im Sinne der Zielsetzung ³⁰³⁾

"Factor analysis was applied for the purpose of determining which of the variables were the most important for prediction ... (It) also provides the means of testing to what extent consumption expenditures are affected by the variants in the family structure"

zu dem Ergebnis führte, daß die fünf strukturellen Variablen die einzelnen Ausgabenarten in ganz unterschiedlichem Maße beeinflussen; der gewöhnlich als "Konsumeinheit" betrachtete Haushalt erfährt eine Differenzierung durch die Möglichkeit, einen quantitativen Zusammenhang zwischen der Aufteilung der Gesamtausgaben nach Ausgabenarten und einigen Merkmalen der Familienstruktur herzustellen. Der Zweck der mit den Methoden der Faktoranalyse vorgenommenen Untersuchung von MILLICAN, "to sort out and establish some of the relationships which exist between certain expenditures

300) MILLICAN, R.D. (1959) - "Application of factor analysis to consumption expenditures", Rev. Econ. Stat. 41, Seite 74-76.

301) Die beiden letzten Variablen wurden nach dem erreichten Ausbildungsstand (level of academic achievement) und nach klimatischen Gegebenheiten (KÖPPEN scale) quantifiziert.

302) Das Quartimax-Verfahren ist eine spezielle Methode zur orthogonalen Bewegung der Faktorachsen auf die Einfachstruktur, das wir im Abschnitt 310 nicht näher behandeln konnten. Vgl. NEUHAUS-WRIGLEY (1954), a.a.O., Seite 81-91.

303) MILLICAN, R.D. (1959), a.a.O., Seite 75.

and the family components" ³⁰⁴⁾, kann auch mit der Komponentenanalyse erreicht werden, wobei durch die Maximierung der Varianzanteile unter Umständen eine stabilere, die Bewegung der Faktorachsen erübrigende Lösung zu erwarten ist. Die von MILLICAN durchgeführte Identifizierung der Faktoren als "personal factor", "age factor", "general factor" und anderen ist für die Untersuchungsergebnisse selbst nicht wesentlich; aber auch dann, wenn man auf eine Faktoranalyse im engeren Sinne Wert legt, ist aus den im Abschnitt 31 ausführlich besprochenen Gründen ³⁰⁵⁾ für die Bewegung von Faktorachsen den Hauptkomponenten gegenüber anderen Faktoren der Vorzug zu geben.

Die in den beiden dargestellten Analysen hervorgehobenen Ergebnisse haben für generelle Betrachtungen der Wirtschaftsrechnungen in privaten Haushalten die folgende Bedeutung. Erhebungen der Verbrauchsstruktur, also der Aufteilung des Gesamtverbrauchs nach Bedarfsgruppen, Ausgabengruppen und weitergehend nach ausgewählten Warenarten, werden in der Regel an Haushaltstypen orientiert, die sich neben der Stellung des Haushaltsvorstands nach der Zahl der Personen und der Höhe der Verbrauchsausgaben gliedern ³⁰⁶⁾. Wenn aber nicht die Verbrauchsstruktur in Abhängigkeit von diesen Merkmalen, sondern die Verbrauchsstruktur schlechthin Gegenstand der Untersuchung sein soll, wäre in Übereinstimmung mit POTI die Frage zu klären, ob sich die gewöhnliche Haushaltsgliederung für diesen speziellen Zweck eignet oder ob nicht eine Aufteilung nach "Strukturtypen" mit Hilfe der Hauptkomponenten dem gewünschten Untersuchungsziel besser entspricht. Die Arbeit von MILLICAN weist insofern in eine ähnliche Richtung, als daß

304) MILLICAN, R.D. (1959), a.a.O., Seite 76.

305) Vgl. insbesondere Seite 121.

306) So in der deutschen amtlichen Statistik der Begriff der Vier-Personen-Arbeitnehmerhaushaltung in der unteren, mittleren oder gehobenen Verbrauchergruppe, siehe zum Beispiel ENGELMANN, M. (1960) - "Der Verbrauch in Arbeitnehmerhaushalten im Jahr 1959", Wirtsch.u.Stat.N.F. 12, Seite 207-211.

wechselseitige Beziehungen zwischen den strukturellen Merkmalen der Haushalte und den Arten der Verbrauchsausgaben hergestellt werden, die gegenüber der sonst angewendeten Gliederung tiefere Einsichten in die Bestimmungsgründe der Verbrauchsstruktur vermitteln. Es erscheint möglich, daß derartige Analysen für die Problematik der statistischen Messung der Veränderungen der Verbrauchsstruktur im Zeitablauf, die gerade jetzt in Verbindung mit der nicht mehr voll den heutigen Gegebenheiten entsprechenden Zusammensetzung des "Warenkorbes 1950" des Preisindex der Lebenshaltung aktuell ist ³⁰⁷⁾, von Interesse sein können. Aber auch bei der sehr schwierigen und bisher nur in Ansätzen vorliegenden Lösung der Probleme des internationalen Vergleichs der Lebenshaltungskosten, dessen Kernpunkt - der Vergleich von "Warenkörben" " ... in einer für jedes der verglichenen Länder typischen Zusammensetzung" ³⁰⁸⁾ - eine Frage der Verbrauchsstruktur ist, sind entsprechende Anwendungen der Komponentenanalyse denkbar.

42 SPEZIELLE ANWENDUNGSGEBIETE

420 Die Verwendung von Hauptkomponenten in der Straßenverkehrsunfallstatistik

In letzter Zeit hat die starke Zunahme der Kraftfahrzeuge in der Bundesrepublik Deutschland ein besonderes Interesse

307) FÜRST, G. (1960) - "Zur Aussagekraft von Preisindexziffern für die Lebenshaltung", Wirtsch. u. Stat. N. F. 12, Seite 5-8.

308) JACOBS, A. (1957) - Der internationale Vergleich der Lebenshaltungskosten. Berlin: Duncker & Humblot, Seite 9. Vgl. ferner
STAEHLE, H. (1934) - International Comparison of Food Costs. In: INTERNATIONAL LABOUR OFFICE (Hrsg.) - International Comparison of Cost of Living (Studies & Reports, Ser. N., No. 20), Seite 1-105.
UNITED NATIONS (Hrsg.) (1954) - Report on International Definition and Measurement of Standards and Levels of Living (Doc. E/CN.3/179; E/CN.5/299).

an jenem Zweig der amtlichen Statistik wachgerufen, der die Erfassung, Darstellung und quantitative Analyse der als Folge der wachsenden Motorisierung auftretenden Straßenverkehrsunfälle umfaßt. Als Zwecke der Straßenverkehrs-unfallstatistik sind neben der laufenden Beobachtung der Entwicklung vor allem die Erforschung der Struktur der Unfälle zu nennen, die das Ziel hat, die zahlreichen als Ursachen vorhandenen Faktoren zu untersuchen und, soweit möglich, zu quantitativen, statistisch einwandfreien Aussagen über ihren Einfluß auf Charakter und Häufigkeit der Unfälle zu gelangen ³⁰⁹⁾. Diese letztlich unter dem Gesichtspunkt der Vermeidung künftiger Straßenverkehrs-unfälle stehende Aufgabenstellung wirft eine Reihe von Problemen auf, unter denen hinsichtlich der genannten Kausalforschung die Frage nach der Art, der Messung und der gegenseitigen Verflechtung der Unfallursachen besonders wichtig ist. Hierzu schreibt HOSSE ³¹⁰⁾:

"Die Faktoren, die das Unfallgeschehen beeinflussen können, sind zahlreich: Verkehrsdichte, Verkehrsstruktur, menschliches Verhalten im Verkehr, Straßen- und Witterungsverhältnisse, technischer Zustand der Fahrzeuge, Bevölkerungsdichte und Zufälle. Es ist nicht möglich, alle diese Faktoren, die sich in ihrer Wirkung teilweise überschneiden, zu quantifizieren, um eine einwandfreie und vollständige Bezugsgröße zu den Unfällen zu erhalten. Bestenfalls lassen sich eine, zwei oder drei Bezugsgrößen ermitteln, die aber niemals den ganzen Komplex erfassen und deshalb nur ungenaue Bezugswerte erbringen."

Trotz der damit angedeuteten Schwierigkeiten ist man versucht, aus dieser vielseitigen Problemstellung heraus eine umfassende Verwendung für die Komponentenanalyse zu entwickeln, jedoch liegt ein unmittelbarer Anlaß unseres Eingehens auf Straßenverkehrsunfälle deswegen vor, weil ein Teilgebiet

309) Vgl. den Abschnitt "Zweck der Straßenverkehrs-unfallstatistik" bei SCHAUB, G. (1958) - "Zur Problematik der deutschen Straßenverkehrs-unfallstatistik. Erwiderung und Ergänzung zum Aufsatz von HOSSE", Allg. Stat. Arch. 42, Seite 29-30.

310) HOSSE, H. (1957) - "Zur Problematik der deutschen Straßenverkehrs-unfallstatistik", Allg. Stat. Arch. 41, Seite 19-20.

bereits mit Hilfe der Komponentenanalyse untersucht worden ist. Dabei handelt es sich um den besonders naheliegenden Ausschnitt einer Analyse der Zusammenhänge zwischen den Kraftfahrzeugbeständen einerseits und gewissen Merkmalen der Straßenverkehrsunfallstatistik andererseits. Wir wollen auf diese von EGERT ³¹¹⁾ stammende Arbeit näher eingehen und im folgenden einige ergänzende Anwendungsmöglichkeiten innerhalb desselben Sachgebiets aufzeigen.

Der Grundgedanke der Untersuchung von EGERT besteht darin, aus der Verteilung der Kraftfahrzeugbestände über die Stadt- und Landkreise des Landes Nordrhein-Westfalen die Hauptkomponenten der Kraftfahrzeugbestandsarten zu berechnen und den entsprechenden regionalen Werten einer größeren Anzahl von Merkmalen der Straßenverkehrsunfallstatistik gegenüberzustellen. Es wird eine Deutung der Hauptkomponenten vorgenommen, jedoch ist diese für den Wert der Ergebnisse nicht als entscheidend anzusehen. Der Zweck der Untersuchung ist demgemäß ein doppelter ³¹²⁾:

"Die im folgenden genannten ... Ergebnisse geben

1. Aufschluß über die den Merkmalen der Straßenverkehrsunfallstatistik zugrunde liegenden Strukturanteile und
2. zeigen sie auf, ob und in welcher Hinsicht sich einzelne Kreise aus der Gesamtheit im Hinblick auf die Straßenverkehrsunfallstatistik unter Berücksichtigung auch ihrer besonderen Struktur herausheben."

Die eigentliche Komponentenanalyse geht von den Beständen für sechs Kraftfahrzeugarten (Krafträder; Personen-, Kombinations- und Krankenkraftwagen; Kraftomnibusse einschließl. Obusse; Lastkraftwagen; zulassungspflichtige Zugmaschinen; Sonderfahrzeuge) in den 95 Kreisen Nordrhein-Westfalens

³¹¹⁾ EGERT, P. (1958) - Kraftfahrzeugbestand und Verkehrsunfälle (Verkehrswiss. Veröffentl. d. Min. f. Wirtsch. u. Verkehr Nordrh.-Westf., Heft 40). Düsseldorf: Droste-Verlag. Der Ausdruck "Komponentenanalyse" wird von EGERT nicht gebraucht, die Methode wird unter Bezugnahme auf die Faktoranalyse (Seite 21) und auf HOTELLING (Seite 68) von Grund auf entwickelt und ausführlich abgeleitet.

³¹²⁾ EGERT, P. (1958), a.a.O., Seite 50.

nach dem Stichtag vom 1. Januar 1956 aus ³¹³⁾ ($n=6; N=95$). Es wurden die drei größten Eigenwerte der Korrelationsmatrix mit zusammen 95,11% der Gesamtvarianz und nach (111') die Hauptkomponenten z_1 , z_2 und z_3 - hier "Hauptvariable" genannt - berechnet ³¹⁴⁾. Die Koeffizienten a_{ik}/λ_k ($i=1,2,\dots,6; k=1,2,3$) des Systems (111') ermöglichen folgende Deutung der Hauptkomponenten: z_1 (70,45% der Gesamtvarianz) ist an allen Variablen mit Ausnahme der Zugmaschinen etwa gleichmäßig beteiligt, sie wird daher als "ein Maß für den Umfang des Verkehrs in den zugeordneten Kreisen" ³¹⁵⁾ aufgefaßt; z_2 (19,92% der Gesamtvarianz) weist für Personenkraftwagen, Omnibusse und Lastkraftwagen positive, für Krafträder und Zugmaschinen negative "Faktorladungen" auf, sie wird dementsprechend als eine Art Skala für den Stadt- (positiv) oder Landcharakter (negativ) eines Kreises angesehen; z_3 mit 4,74% der Gesamtvarianz läßt keine Deutung mehr zu.

An die Bestimmung der Hauptkomponenten schließt EGERT die bereits angeführte Prüfung der Zusammenhänge zwischen den z_k ($k=1,2,3$) und 33 Merkmalen aus der Straßenverkehrsunfallstatistik an, die nach Jahresergebnissen 1953, 1954 und 1955 sowie nach Unfallbeteiligten und vorläufigen Unfallursachen (je 15 Reihen, dazu die Gesamtzahlen der Unfälle, der Verletzten und der Getöteten Personen) gegliedert sind. Diese Untersuchungsstufe bedient sich der linearen Regressionsanalyse, deren Anwendung wie folgt gekennzeichnet wird ³¹⁶⁾:

313) Die Zahlen stammen aus MINISTERIUM FÜR WIRTSCHAFT UND VERKEHR, NORDRHEIN-WESTFALEN(Hrsg.)(1956) - Die Verkehrsentwicklung des Landes Nordrhein-Westfalen, Nr. 27: Jahresbericht 1955, Tabelle 10, Seite 27* - 28*. Sämtliche Werte wurden durch Logarithmierung mit den Unfallzahlen, die nach EGERT(1958, a.a.O., Seite 37-40) in guter Annäherung einer logarithmischen Normalverteilung folgen, in Übereinstimmung gebracht.

314) EGERT,P.(1958), a.a.O., Gleichung (56), Seite 45 und Tabelle VI, Seite 46-47.

315) EGERT,P.(1958), a.a.O., Seite 43.

316) EGERT,P.(1958), a.a.O., Seite 50.

"Die geeigneten Maßgrößen, die den Strukturanteil eines Merkmals ausdrücken, sind die Regressionskoeffizienten nach den Variablen Da diese ... voneinander unabhängig sind, sind die nach ihnen gebildeten Koeffizienten unabhängige Aussagen. Ist ein solcher Koeffizient groß und positiv, so ist das betreffende Merkmal stark im gleichlaufenden Sinne von der betreffenden Strukturvariablen abhängig. Ist der Koeffizient fast Null oder negativ, so liegt Unabhängigkeit oder gegenläufige Tendenz vor."

Aus Raumgründen müssen wir uns hier auf die Wiedergabe von neun der insgesamt berechneten 99 Regressionskoeffizienten beschränken. Wir wählen das Merkmal "Getötete insgesamt", dessen Koeffizienten auf Landesebene folgende Werte annehmen ³¹⁷⁾.

Tabelle 19: Koeffizienten der linearen Regressionsgleichungen zwischen den ersten drei Hauptkomponenten und der Gesamtzahl der bei Straßenverkehrsunfällen getöteten Personen in Nordrhein-Westfalen, 1953-1955

Hauptkomponente	Regressionskoeffizienten		
	1953	1954	1955
(1)	(2)	(3)	(4)
z_1	2,552	2,854	2,840
z_2	-0,741	-0,740	-0,679
z_3	0,309	0,515	0,206

Der ausführlichen Deutung, die von EGERT für diese und alle anderen Koeffizienten vorgenommen wird, entnehmen wir als bemerkenswertes Ergebnis der Tabelle 19, daß die Zahl der Unfalltoten zwar mit wachsendem "Verkehrsvolumen" (z_1) zunimmt, mit steigender "Verkehrskompaktheit" (z_2) aber erheblich fällt, wobei "Verkehrskompaktheit" eine begriffliche Fassung der durch die zweite Hauptkomponente quantifizierten Unterscheidung der Kreise nach ihrem städtischen oder ländlichen Charakter ist ³¹⁸⁾. Der eigentliche

³¹⁷⁾ EGERT, P. (1958), a.a.O., Tabelle VII auf Seite 51.

³¹⁸⁾ Zur Abnahme der Getötetenzahlen bei Zunahme von z_2 schreibt EGERT (1958, a.a.O., Seite 51): "Ich möchte auf diesen Umstand mit allem Nachdruck verweisen, da die überwiegende Zahl aller Äußerungen zum Thema Verkehrsunfälle gerade von der Voraussetzung eines gegenteiligen Verhaltens ausgeht."

Kreisvergleich als zweiter Punkt der Aufgabenstellung ³¹⁹⁾ beruht auf einer Methode, die es ermöglicht, für jeden Kreis und für jedes Merkmal die Differenz zwischen dem durch die Werte z_{kj} ($k=1,2,3; j=1,2,\dots,95$) der drei Hauptkomponenten gegebenen "Normalwert" und dem dafür tatsächlich beobachteten Wert zu berechnen. Die Größe dieser Differenzen, von denen für 21 Merkmale und 95 Kreise insgesamt 1995 ermittelt wurden ³²⁰⁾, dient als Maß dafür, inwieweit sich einzelne Kreise hinsichtlich bestimmter Merkmale der Straßenverkehrsunfallstatistik von der durch die strukturellen Variablen z_k implizierten "Normalsituation", bei der die Differenz verschwindet, entfernen und kann als direkter Hinweis auf spezielle Unfallverhütungsmaßnahmen betrachtet werden. Allerdings muß man allen Einzelheiten der Untersuchung von EGERT entgegenhalten, daß für Vergleichszwecke mit regionalen Unfallzahlen die Bezugnahme auf die Kraftfahrzeugbestände in den Kreisen unbefriedigend bleibt, weil zwischen beiden Merkmalsgruppen keine eindeutige Zuordnung besteht. Man denke nur daran, daß namentlich in Großstädten - aber zum Beispiel auch auf Fernstraßen, die durch Landkreise führen - an Unfällen vielfach Fahrzeuge beteiligt sind, die ihren Standort in einem anderen Kreis, einem anderen Bundesland oder gar im Ausland haben.

Das von EGERT zugrunde gelegte System von Ausgangsvariablen kann durch Hinzufügung weiterer Kreismerkmale, die für die Verkehrsstruktur relevant sind, ausgedehnt werden; diese können zum Beispiel in einer stärkeren Untergliederung der Kraftfahrzeugbestände bestehen oder gewisse Informationen über die Straßenverhältnisse ³²¹⁾ (Aufteilung nach Bundesautobahnen, Bundesstraßen, Landstraßen I. und II. Ordnung,

319) EGERT, P. (1958), a.a.O., Seiten 33 und 58. Vgl. auch Seite 166.

320) EGERT, P. (1958), a.a.O., Tabelle XII neben Seite 68.

321) STATISTISCHES LANDESAMT NORDRHEIN-WESTFALEN (Hrsg.) (1959) - Die Gemeindestraßen in Nordrhein-Westfalen am 31. März 1956 (Beitr.z.Stat.Nordrh.-Westf., Heft 108).

Inner- und Außenortstraßen) enthalten. Dadurch ist es möglich, aus der "Faktorstruktur" zusätzliche Aufschlüsse zu gewinnen und die Gegenüberstellung mit den Unfallmerkmalen zu verfeinern ³²²⁾. Andererseits kann die Komponentenanalyse in diesem Zusammenhang auch dann angewendet werden, wenn nicht N Regionen, sondern N Zeiteinheiten als Gliederungsgesichtspunkte gewählt werden, wie es etwa bei Untersuchungen der Witterungseinflüsse auf das Unfallgeschehen erforderlich ist ³²³⁾.

421 Komponentenanalyse und statistische Regionalforschung

Unter den wirtschafts- und sozialwissenschaftlichen Anwendungen der Komponentenanalyse sind häufig solche zu finden, deren Ausgangsgesamtheit die Verteilung von n statistischen Merkmalen über N regionale Einheiten enthält. Dies ist zum Beispiel bei der Ermittlung der landwirtschaftlichen Produktivität durch KENDALL und in der Untersuchung der Verkehrsstruktur von EGERT der Fall gewesen ³²⁴⁾. Während aber dort die regionale Aufteilung nur als eine unter verschiedenen möglichen Gliederungsgesichtspunkten auf dem Wege zur Gewinnung von Einsichten, die entweder überhaupt nicht oder nur mittelbar regional bezogen sind, aufzufassen war, gibt

322) Auch diese können tiefer gegliedert werden, siehe zum Beispiel den Abschnitt "Kreisergebnisse" in STATISTISCHES LANDESAMT NORDRHEIN-WESTFALEN (Hrsg.) (1960) - Straßenverkehrsunfälle in Nordrhein-Westfalen 1956-1958 (Beitr.z.Stat.Nordrh.-Westf., Heft 119), Seite 50-61. Weitere regionale Merkmale, vor allem gewisse Beziehungszahlen, sind bei KLEINE, W. (1960) - "Straßen-, Kraftfahrzeug- und Unfall-dichte in Nordrhein-Westfalen im Jahre 1958", Stat.Rundsch.Nordrh.-Westf. 12, Seite 32-36, zu finden.

323) KLEINE, W. (1959) - "Wetter und Straßenverkehrsunfälle in Nordrhein-Westfalen", Stat.Rundsch.Nordrh.-Westf. 11, Seite 217-223.

324) Vgl. die Abschnitte 412, Seite 140-145, und 420, Seite 166-170.

es weitere Anwendungsmöglichkeiten, bei denen das Konzept des Raumes im Vordergrund steht. Wir gehen damit zu Problemen über, die die Bedeutung der Komponentenanalyse für die statistische Regionalforschung betreffen.

Eine wichtige Aufgabe der Raumforschung in ihrer Erscheinungsform als wissenschaftlicher Methode zur statistischen Untersuchung der wirtschaftlichen und sozialen Gegebenheiten des Raumes ³²⁵⁾ ist die als Voraussetzung zur regionalen Differenzierung und Analyse erforderliche Bildung von Raumeinheiten. Dieses Problem tritt wegen der grundsätzlich unbegrenzt großen Zahl von Möglichkeiten zur Abgrenzung von Raumeinheiten in sehr vielen Variationen auf, jedoch lassen sich als übergeordnete Einteilungskriterien zwei ihrem Wesen nach gegensätzliche, wenn auch gleichzeitig anwendbare Fälle unterscheiden, nämlich die homogenen und die funktionalen Raumeinheiten. Erstere werden "durch die Gleichartigkeit oder Ähnlichkeit eines oder mehrerer Strukturmerkmale" gekennzeichnet, während funktionale Räume als "durch eine mehr oder weniger große Zahl überörtlicher Wechselbeziehungen verschiedenster Art und deren Orientierung auf ein wirtschaftliches Gravitationszentrum zusammengeschlossen" zu betrachten sind ³²⁶⁾.

- 325) Zur Begriffsbildung der Raumforschung und ihrer Teilgebiete, insbesondere der regionalen Wirtschaftsforschung, sowie zur Stellung der Statistik als regionaler Forschungsmethode vgl. BOUSTEDT, O. (1955) - "Statistical methods and problems in regional research", Bull. Inst. Int. Stat. 35, 4ème Livr., Seite 357-375.
ESENWEIN-ROTHE, I. (1955) - "Statistik als Methode der Raumforschung", Allg. Stat. Arch. 39, Seite 136-143.
BOUSTEDT, O.; RANZ, H. (1957) - Regionale Struktur- und Wirtschaftsforschung. Bremen: W. Dorn-Verlag, Seite 9-30.
- 326) BOUSTEDT-RANZ (1957), a.a.O., Seite 36-37. Siehe auch BOGUE, D. J. (1955) - "Nodal versus homogeneous regions and statistical techniques for measuring the influence of each", Bull. Inst. Int. Stat. 35, 4ème Livr., Seite 377-397. Ein Beispiel für die homogene Raumgliederung ist die in der Bundesrepublik amtlicherseits aufgestellte "Naturräumliche Gliederung Deutschlands", dagegen ist das deutsche System der Verkehrsbezirke funktional orientiert. Vgl. FÜRST, G. (1955) - "Gedanken zur regionalen Gliederung in der Wirtschaftsstatistik", Wirtsch. u. Stat. N.F. 7, Seite 182-183.

Ein statistisches Verfahren zur Abgrenzung von homogenen Raumeinheiten ist die "Feste Index-Methode" (fixed index method), deren Wesen BOGUE wie folgt beschreibt ³²⁷⁾:

"A few selected indexes are given a series of statistical weights or rankings of relative importance (they may have equal weights). The indexes are computed for each of the parcels of land to be classified. Parcels with similar index values are classified together as a group. Then an effort is made to enclose within one set of boundaries as many of the parcels in each group as possible, while including as few parcels as possible from adjacent groups."

Wir wollen nun zeigen, wie die Komponentenanalyse - hierbei stets als spezielles Verfahren der Faktoranalyse aufgefaßt - in einigen Fällen zur quantitativ exakten Durchführung der "Festen Index-Methode" herangezogen wurde.

Eine Studie über die Aufstellung von agrarsoziologisch homogenen Raumeinheiten für den amerikanischen Bundesstaat Ohio bildete die Grundlage einer Untersuchung von HAGOOD, DANILEVSKY und BEUM, in der diese Form der Anwendung der Komponentenanalyse vermutlich zum ersten Male geprüft worden ist ³²⁸⁾. Das Problem der Bildung von homogenen Raumeinheiten wird in die drei Stufen Wahl der zugrunde liegenden Regionen, Bestimmung der sozialökonomischen Strukturmerkmale ³²⁹⁾ und Synthese der ausgewählten Merkmale unter dem Gesichtspunkt maximaler Homogenität zerlegt; die Komponentenanalyse findet in der dritten Stufe eine Verwendung,

327) BOGUE, D.J. (1955), a.a.O., Seite 382-383.

328) HAGOOD, M.J.; DANILEVSKY, N.; BEUM, C.O. (1941) - "An examination of the use of factor analysis in the problem of subregional delineation", Rur.Sociol. 6, Seite 216-233.

329) Diese umfaßten nach dem Stande von 1930 je 111 Daten, zum Beispiel Einkommen landwirtschaftlicher Betriebe (farm income) und Wohlstandsindex der Landbevölkerung (rural plane of living index), für 88 Kreise (counties). Vgl. LIVERY, C.E.; ALMACK, R.B. (1938) - A Method of Determining Rural Social Sub-Areas with Application to Ohio (Ohio State Univ. and Ohio Agr.Exp.Stat., Dept. Rur.Econ.Bull.No. 106). (Angaben nach HAGOOD-DANILEVSKY-BEUM, 1941, a.a.O., Seiten 218 und 222).

die methodisch der Kombination der normierten Strukturmerkmale durch die erste Hauptkomponente entspricht ³³⁰⁾. Die Größen z_{1j} ($j=1,2,\dots,88$) stellen eine Folge von Indexwerten dar, die einen maximalen Anteil an der Gesamtvariation der Strukturmerkmale aufweisen; ihre Zusammenfassung zu Größenklassen bringt daher alle die Kreise zueinander, deren Homogenität in Bezug auf die x_i ($i=1,2,\dots,111$) maximal ist. Interessant ist, daß man den erwünschten geographischen Zusammenhang der als homogen klassifizierten Kreise dadurch zu erreichen versucht hat, daß als zusätzliches Merkmal die Längen- und Breitenkoordinaten des (flächenmäßigen) Kreiscentrums - einheitlich auf einen willkürlichen Punkt der Karte von Ohio bezogen - eingeführt wurden. Allerdings ist dieser " 'pulling together' effect of the distance characteristics" ³³¹⁾ wegen der Überzahl anderer Merkmale relativ gering, so daß eine Reihe von Kreisen keine gemeinsamen Grenzen mit den übrigen Kreisen derselben Strukturgruppe besitzt und sich Kompromisse in der Aufstellung geographisch abgegrenzter Raumeinheiten nicht vermeiden lassen.

In methodischer und sachlicher Hinsicht erfuhr die Verwendung der Komponentenanalyse in einer weiteren Arbeit ³³²⁾ von HAGOOD eine beachtliche Ausweitung. Die Aufgabenstellung bestand hierbei darin, das gesamte Gebiet der USA

330) HAGOOD-DANILEVSKY-BEUM(1941), a.a.O., Gleichung (1) auf Seite 222. Wir können hier auf die Wiedergabe von Formeln verzichten, weil dieses Vorgehen mit dem von (111), Seite 53, (186), Seite 142 und (188), Seite 157, für den Fall ($k=1; n=111; N=88$) identisch ist. Als methodische Grundlage verwenden die Autoren die Arbeit von WILKS, S.S. (1938) - "Weighting systems for linear functions of correlated variables when there is no independent variable", Psychometrika 3, Seite 23-43, in der die Eigenschaften der Linearkombination (111) speziell für ($k=1$) untersucht werden.

331) HAGOOD-DANILEVSKY-BEUM(1941), a.a.O., Seite 228.

332) HAGOOD, M.J. (1943) - "Statistical methods for delineation of regions applied to data on agriculture and population", Social Forces 21, Seite 287-297. In Kurzfassung wiedergegeben bei HAGOOD, M.J.; PRICE, D.O. (1952 ²) - Statistics for Sociologists. New York: H. Holt & Co., Seite 541-546.

in eine Reihe von größeren Regionen aufzuteilen, die aus aneinander grenzenden Einzelstaaten bestehen und hinsichtlich einer großen Zahl von Merkmalen aus der Bevölkerungs- und Landwirtschaftsstatistik nach den Zählungen von 1940 möglichst homogen sein sollten. Die Strukturmerkmale waren in je 52 bevölkerungs- und agrarstatistische Reihen nach Maßgabe der folgenden Übersicht gegliedert.³³³⁾

Tabelle 20: Gliederung von 104 bevölkerungs- und landwirtschaftsstatistischen Merkmalen in Merkmalsgruppen. Angaben nach HAGOOD.

Bevölkerungsstatistik		Landwirtschaftsstatistik	
Merkmalgruppe (mit Kurzbezeichnung)	Zahl der Merkmale	Merkmalgruppe (mit Kurzbezeichnung)	Zahl der Merkmale
(1)	(2)	(3)	(4)
Wohnungsverhältnisse (Residence) Pa	6	Bodenbenutzung (Land Use) Aa	4
Rassische Zusammensetzung (Race) Pb	5	Erntestatistik (Crops) Ab	12
Geschlechterverteilung (Sex) Pc	6	Viehhaltung (Livestock) Ac	10
Altersverteilung (Age) Pd	6	Landwirtschaftliche Besitzverhältnisse (Farm Tenure) Ad	11
Bildungsstand (Education) Pe	5	Vermögenswerte der landwirtschaftlichen Betriebe (Farm Values) Ae	8
Arbeitsmarkt (Employment Status) Pf	9	Finanzstatus der landwirtschaftlichen Betriebe (Farm Finance) Af	7
Erwerbstätigkeit (Occupation) Pg	8		
Bevölkerungsbewegung (Vital Statistics) Ph	7		
Insgesamt	52	Insgesamt	52

Zunächst wurden die 14 Merkmalsgruppen nach dem beschriebenen Verfahren einzeln analysiert ($n=6,5,6, \dots$; $N=48$) und zur Bildung von je vier homogenen Regionen mit etwa 12 Staaten benutzt, wobei der geographische Zusammenhang

³³³⁾ Siehe Seite 175.

nicht in jedem Falle voll gewahrt blieb ³³⁴⁾. Auf einer zweiten Stufe erhielt man durch die Berechnung der ersten Hauptkomponenten aus der Korrelationsmatrix der 14 Indexreihen der ersten Stufe ($n=14; N=48$) eine Gleichung für den "Gesamtindex" (composite index) der "agriculture-population profiles" ³³⁵⁾ der Einzelstaaten, die für normierte Größen unter Verwendung der Kurzbezeichnungen der Tabelle 20 die Form

$$\begin{aligned}
 (190) \quad z_{1j} &= \frac{1}{\lambda_1} \sum_{i=1}^{14} a_{ik} x_{ij} \\
 &= 0,308x_{Aa,j} + 0,728x_{Ab,j} - 0,924x_{Ac,j} + 0,923x_{Ad,j} \\
 &\quad - 0,971x_{Ae,j} - 0,943x_{Af,j} - 0,504x_{Pa,j} + 0,866x_{Pb,j} \\
 &\quad - 0,455x_{Pc,j} + 0,958x_{Pd,j} + 1,000x_{Pe,j} - 0,116x_{Pf,j} \\
 &\quad - 0,926x_{Pg,j} + 0,902x_{Ph,j} \quad (j=1,2,\dots,48)
 \end{aligned}$$

333) HAGOOD-PRICE(1952), a.a.O., Seite 541-542, fehlerhafte Angaben für die Bevölkerungsstatistik berichtigt nach HAGOOD, M.J.(1943), a.a.O., Seite 295. Der letztgenannten Veröffentlichung, Seiten 289 und 292, entnehmen wir als Beispiele für die Zusammensetzung der Merkmalsgruppen die nachstehenden Reihen: "Educational Status (persons 25 years of age and older), 1 = Median years of school completed, 2 = Percent with less than 5 years of school completed, 3 = Ratio of those completing 4 years or more of college to all with some high school or more, 4 = Ratio of median years of school completed for urban population to that for rural population, 5 = Ratio of percent completing 4 or more years of college for urban population to that for rural population."
 "Land Use, 1 = Percent farm land of all land, 2 = Percent cropland (harvested and failure) of all farm land, 3 = Percent woodland of farm land not used for crops, 4 = Mean size of farms (acres)."

334) Vgl. zum Beispiel die Karte der USA bei HAGOOD-PRICE(1952), a.a.O., Seite 543, die die Gliederung der Staaten nach dem "Educational Status Index" zeigt. Die z_{1j} ($j=1,2,\dots,48$) wurden so transformiert, daß sie als niedrigsten Wert 100, als höchsten 0 ergaben; die Klassenintervalle waren 0-13, 14-22, 23-44 und 44-100.

335) HAGOOD-PRICE(1952), a.a.O., Seiten 542 und 544.

annahm. Durch die Aufstellung von Klassenintervallen für die 48 Werte von (190) war die Bildung von sechs Raumeinheiten möglich, die hinsichtlich aller 104 Ausgangsreihen größtmögliche Homogenität aufwiesen. Allerdings gab es auch hier Schwierigkeiten mit einigen Staaten, die gewissermaßen als "Außenseiter" den Indexwerten ihrer Nachbarstaaten nicht genügend entsprachen; für die geographisch zusammenhängende Einteilung mußten daher durch die Berechnung von "Ähnlichkeitskoeffizienten" (coefficients of similarity) zusätzliche Entscheidungen vorgenommen werden ³³⁶).

Weitere Anwendungen der Komponentenanalyse unter dem Gesichtspunkt der statistischen Regionalforschung, auf deren Erwähnung wir uns hier beschränken müssen, hat das United States Bureau of Agricultural Economics bei der Konstruktion eines Wohlstandsindex der Landbevölkerung (rural level-of-living index) ³³⁷) für 3000 Kreise und bei der Bildung von homogenen Raumeinheiten für die Zwecke des geschichteten Stichprobenverfahrens durchgeführt ³³⁸). Auf einer etwas anderen Ebene liegt eine Arbeit von PRICE ³³⁹), in der die Werte von 15 sozialökonomischen Variablen für die 93 amerikanischen Großstädte, die 1930 eine Bevölkerung von 100 000 oder mehr Einwohnern hatten, einer Faktoranalyse

336) HAGOOD, M.J. (1943), a.a.O., Seite 293.

HAGOOD-PRICE (1952), a.a.O., Seite 542-546.

337) HAGOOD-PRICE (1952), a.a.O., Seite 526-530. Es handelte sich um vier Ausgangsvariable (1 = Percentage of farms with electricity in farm dwelling, 1945, 2 = Percentage of farms with telephone in farm dwelling, 1945, 3 = Percentage of farms with automobiles, 1945, 4 = Mean value of products sold or traded per farms reporting, 1944, in hundreds of dollars) und 196 Kreise (n=4; N=196), die auf Stichprobenbasis aus der Gesamtheit der 3000 Kreise ausgewählt wurden.

338) HAGOOD, M.J.; BERNERT, E.H. (1945) - "Component indexes as a basis for stratification in sampling", J. Am. Stat. Ass. 40, Seite 330-341. Abgedruckt bei HAGOOD-PRICE (1952), a.a.O., Seite 530-541.

339) PRICE, D.O. (1942) - "Factor analysis in the study of metropolitan centers", Social Forces 20, Seite 449-455.

(Schwerpunktmethode) mit ($n=15; N=99$) unterworfen wurden. Die als signifikant angesehenen ersten vier Faktoren konnten auf Grund der Faktorstruktur mit den Bedeutungen "Großstadtcharakter" (factor of maturity), "Handels- und Verkehrszentrum" (service center, lack of predominance of industry), "Allgemeiner Lebensstandard" (general level of living) und "Handelsumsätze je Kopf der Bevölkerung" (per capita trade volume) versehen werden. Diesen Identifizierungen sind die üblichen Vorbehalte entgegenzustellen, jedoch möchten wir uns unter Betonung der besseren Eignung der Komponentenanalyse der insbesondere auf die Soziometrie bezogenen Feststellung ³⁴⁰⁾

"In spite of the fact that the invariance of the four factors obtained has not been investigated, factor analysis seems to present itself as a powerful new tool in sociological research. If the sociologist is attempting to get some orderly picture of society, he must, as far as possible, locate the fundamental factors in society from which the other characteristics can be predicted, and factor analysis seems to be a method for moving in that direction."

anschließen.

Die vorgetragenen Anwendungsbeispiele in der statistischen Regionalforschung berühren nach unserer Ansicht nur einen Teil der für die Methode der Hauptkomponenten in Frage kommenden Einsatzmöglichkeiten. Zum Abschluß dieses Abschnitts seien einige Probleme aufgeführt, die für entsprechende Spezialuntersuchungen geeignet sind. So erscheint es zweckmäßig, auch bei der Bildung der bereits erwähnten funktionalen Raumeinheiten, besonders der Wirtschaftsräume, neben den sonst benutzten statistischen Methoden ³⁴¹⁾ auch die Komponentenanalyse in Erwägung zu ziehen. Ferner kann in diesem Zusammenhang an das Problem der Typisierung gedacht

340) PRICE, D.O. (1942), a.a.O., Seite 455. Zu den hier angesprochenen Fragen der statistischen Erforschung von Stadtregionen vgl. auch DENEFFE, P. (1949) - "Soziale Topographie der Großstadt", Allg.Stat.Arch. 33, Seite 444-471.

341) Siehe zum Beispiel VINING, R. (1953) - "Delimitation of economic areas: Statistical conceptions of the spatial structure of an economic system", J.Am.Stat.Ass. 48, Seite 44-64.

werden, das methodisch mit dem der Bildung von homogenen Raumeinheiten verwandt ist und in der Regionalstatistik vor allem auf der Ebene der Gemeindetypisierung auftritt ³⁴²⁾ Schließlich sei angedeutet, daß im Bereich der regionalen Wirtschaftsforschung die Aufgabe der Berechnung von wirtschaftlichen Leistungswerten, zum Beispiel Kaufkraftziffern, für kleinere Gebietseinheiten ³⁴³⁾ - in der Regel kreisfreie Städte und Landkreise - mindestens hinsichtlich einer durch die Hauptkomponenten erzielbaren Verringerung der Dimension einer großen Anzahl von Gebiet zu Gebiet variierender Indikatoren der Wirtschaftskraft ein interessantes Anwendungsgebiet für die Komponentenanalyse sein dürfte.

422 Anwendungen in der technischen Statistik; Ausblick auf verwandte Probleme

Innerhalb der technischen Statistik fällt der statistischen Qualitätskontrolle die Aufgabe zu, Methoden für die Überprüfung der Güte eines Produktes während des Fabrikationsprozesses bereitzustellen. Meistens wird diese laufende Fertigungsüberwachung durch Kontrollkarten vorgenommen, mit denen die Ergebnisse des Produktionsvorgangs in der verschiedensten Weise in zeitlichen Abständen festgehalten und auf mathematisch-statistischer Grundlage hinsichtlich ihres Verhaltens in Bezug auf vorgeschriebene Toleranzen überprüft werden ³⁴⁴⁾. Oft beschränkt sich die Kontrolle

- 342) BOUSTEDT, O. (1951) - "Die statistischen Probleme der sozial-ökonomischen Topographie", Allg. Stat. Arch. 35, Seite 291-310.
 LINDE, H. (1952) - "Grundfragen der Gemeindetypisierung", Forschungs- u. Sitzungsber. d. Akad. f. Raumf. u. Landesplg. 3, Seite 58-121.
 BOUSTEDT-RANZ (1957), a.a.O., Seite 102-105.
- 343) GRIESMEIER, J. (1955) - "Berechnung wirtschaftlicher Leistungswerte für kleinere Gebietseinheiten", Allg. Stat. Arch. 39, Seite 17-24.
 WIRNSHOFER, J. (1958) - "Berechnung wirtschaftlicher Leistungswerte für kreisfreie Städte und Landkreise", Zeitschr. d. Bayer. Stat. Landesamts 90, Seite 1-13.
- 344) Siehe zum Beispiel
 SCHAAFSMA, A. H.; WILLEMZE, F. G. (1959²) - Moderne Qualitätskontrolle. Eindhoven: Philips Technische Bibliothek, Seite 121-189.

auf die Beobachtung eines einzigen Merkmals des in der Fertigung befindlichen Produkts. Es kommt jedoch vor, daß die Güte eines Erzeugnisses von mehreren Merkmalen in einer Weise gemeinsam bestimmt wird, die die getrennte, wenn auch gleichzeitige Beobachtung der einzelnen Variablen ausschließt ³⁴⁵⁾:

"Of increasing importance in the field of statistical quality control is the simultaneous control of two or more related variables when the quality depends on the joint effect of these variables rather than on the effect of each variable separately."

Für Probleme dieser Art bestehen Methoden, die sich unter dem Begriff der mehrdimensionalen Qualitätskontrolle (multivariate quality control) vereinigen lassen und zu einem großen Teil auf HOTELLING ³⁴⁶⁾ zurückgehen. Sie beruhen auf der Theorie der Stichproben aus einer mehrdimensionalen Normalverteilung und für Kontrollkartenzwecke auf einer von HOTELLING entwickelten Prüfgröße T^2 , die eine Verallgemeinerung der t-Verteilung auf den mehrdimensionalen Fall darstellt ³⁴⁷⁾. Hierbei bestehen zunächst keine Anwendungsmöglichkeiten für die Komponentenanalyse ³⁴⁸⁾,

345) JACKSON, J.E. (1956) - "Quality control methods for two related variables", Ind.Qual.Contr. 12, No.7, Seite 4.

346) HOTELLING, H. (1947) - Multivariate Quality Control, Illustrated by the Air Testing of Sample Bombsights. In: EISENHART, C. (Hrsg.) - Selected Techniques of Statistical Analysis for Scientific and Industrial Research and Production and Management Engineering. New York (London): McGraw Hill Book Co., Seite 113-184.

347) HOTELLING, H. (1931) - "The generalization of STUDENT'S ratio", Ann.Math.Stat. 2, Seite 360-378. Vgl. auch CRAMÉR, H. (1946), a.a.O., Seite 403-410.

348) Mit Ausnahme einer gewissen Analogie zum Fall ($n=p=2$), wie er von JACKSON (1956, a.a.O., Seite 5-7) im Zusammenhang mit den Hauptachsen einer Konzentrationsellipse, die als graphische Alternative zu einer T^2 -Kontrollkarte verwendet werden kann, demonstriert und in einer neueren Arbeit als direkte Fragestellung der Komponentenanalyse auf höhere Dimensionen verallgemeinert worden ist. Vgl. Abschnitt 172, Seite 55-58, und JACKSON, J.E. (1959) - "Quality control methods for several related variables", Technometrics 1, Seite 359-377.

jedoch sind in einer aus der Qualitätskontrolle des Entwicklungsprozesses von Farbfilmen stammenden Untersuchung von JACKSON und MORRIS ³⁴⁹⁾ Schwierigkeiten bei der Berechnung von T^2 aufgetreten, die sich durch die Einführung von Hauptkomponenten beseitigen ließen. Wir wollen dazu annehmen, daß n Variable auf den Entwicklungsprozeß einwirken und für die laufende Überwachung unter Beobachtung gehalten werden. Für die Bestimmung von T^2 ist die Inversion der Kovarianzmatrix der n Variablen erforderlich. Wenn diese untereinander sehr hohe Korrelationen aufweisen, führt die Inversion numerisch zu ungenauen Ergebnissen ³⁵⁰⁾. Der Ausweg besteht nun darin, an die Stelle der n Ausgangsvariablen eine im Sinne der Variationserfassung ausreichende Zahl ihrer Hauptkomponenten zu setzen. Diese sind linear unabhängig und aus den Ausgangsvariablen linear transformiert, sie bilden daher ebenfalls eine mehrdimensionale Normalverteilung, für die erneut ein T^2 berechnet werden kann. JACKSON und MORRIS haben dieses Vorgehen an einem Beispiel von sechs Kontrollvariablen, die über einen Zeitraum von 31 Fertigungswochen beobachtet wurden, ausführlich dargestellt ³⁵¹⁾. Es müßte jedoch in einer weiteren Untersuchung geklärt werden, ob nicht im Falle ($p < n$), das heißt dann, wenn sich unter den Ausgangsvariablen linear abhängige befinden, die Transformation auf die Hauptkomponenten ohnehin die einzige Anwendungsmöglichkeit für die Größe T^2 bildet, weil die letztere Stichproben aus einer nichtsingulären mehrdimensionalen Normalverteilung voraussetzt ³⁵²⁾. Ganz abgesehen von den Fragen der mehrdimensionalen Qualitätskontrolle könnte man auch daran denken,

349) JACKSON, J.E.; MORRIS, R.H. (1957) - "An application of multivariate quality control to photographic processing", J.Am.Stat.Ass. 52, Seite 186-199.

350) Zur Gleichung von T^2 siehe zum Beispiel CRAMÉR, H. (1946), a.a.O., (29.11.3) auf Seite 409. Die Schwierigkeit bei der Inversion entsteht dadurch, daß die Kovarianzmatrix bei hohen Korrelationen zwischen den Variablen nahezu singulär ($\text{Rang} < n$) wird. Dann ist aber auch ihre Determinante, die man für die Berechnung der invertierten Matrix benötigt, nahe Null.

351) JACKSON-MORRIS (1957), a.a.O., Seite 190-198.

352) CRAMÉR, H. (1946), a.a.O., Seite 403.

die erste Hauptkomponente einer größeren Anzahl von qualitätsbeeinflussenden Merkmalen zu bestimmen und auf diese, genügende Erfassung der Gesamtvariation vorausgesetzt, die gewöhnlichen Kontrollkartenverfahren für ein Merkmal anzuwenden. Sofern nicht eine hohe Korrelation zwischen z_1 und einem der Ausgangsmerkmale besteht, müßte man dazu allerdings eine ständige Umrechnung nach (111) für ($k=1$) vornehmen, da die Werte nicht direkt beobachtet werden können.

Weitere Anwendungen der Komponentenanalyse in der statistischen Qualitätskontrolle sind nach unserem Wissen bisher nicht vorgenommen worden. Jedoch liegen einige Veröffentlichungen vor, die insoweit der technischen Statistik angehören, als sie die Untersuchung der Struktur gewisser Produkte mit Hilfe der Faktoranalyse betreffen. So haben HARPER, KENT und BLAIR zehn technische Messungen an Kunststoffen mit der Schwerpunktmethode analysiert ³⁵³⁾. Andere Arbeiten beschäftigen sich mit der faktoranalytischen Überprüfung der Eigenschaften von Nahrungsmitteln, insbesondere von Käsesorten ³⁵⁴⁾. Wir wollen zum Abschluß auf einige weitere Probleme hinweisen, die aus der Anwendung der Faktoranalyse auf medizinische und anthropometrische Fragestellungen erwachsen sind und auch für industrielle Zwecke, namentlich solche der Bekleidungsindustrie, eine gewisse Bedeutung besitzen. Dabei handelt es sich um die Analyse einer größeren Zahl von physischen Maßen, wie sie etwa von BURT ³⁵⁵⁾ an einer Gruppe von 2400 Angehörigen

353) HARPER, R.; KENT, A. J.; BLAIR, G. W. S. (1950) - "The application of multiple factor analysis to industrial test data", Brit. J. Appl. Phys. 1, Seite 23-28.

354) HARPER, R.; BARON, M. (1948) - "Factorial analysis of the rheological properties of cheese", Nature 162, Seite 821.

HARPER, R. (1956) - "Factor analysis as a technique for examining complex data on foodstuffs", Appl. Stat. 5, Seite 32-48.

355) BURT, C. (1947) - "Factor analysis and physical types", Psychometrika 12, Seite 171-188. Vgl. ferner HAMMOND, W. H. (1942) - "An application of BURT'S multiple general factor analysis to the delineation of physical types", Man 42, Seite 4-11.

BURT, C. (1944) - "The factorial study of physical types", Man 44, Seite 82-86.

BANKS, C.; BURT, C. (1947) - "Factor analysis of body measurements for British adult males", Ann. Eugen. 13, Seite 238-256.

der Royal Air Force durchgeführt worden ist. Obwohl hierbei Fragen der Herausbildung von Typen nach der regionalen Herkunft und altersmäßigen Zusammensetzung im Vordergrund standen, bieten sich weitere Verwendungsmöglichkeiten an, die die Anpassung des Sortiments an Bekleidungsgegenständen der verschiedensten Art sowohl in Bezug auf die Größenverhältnisse als solche als auch hinsichtlich der mengenmäßigen Verteilung auf die einzelnen Größenklassen zum Gegenstand haben. Eine entsprechende Untersuchung von 29 Körpermaßen an 4128 amerikanischen Frauen stammt von HEATH, eine Reihe von Anregungen für weitere Arbeiten auf diesem Gebiet kann den Veröffentlichungen von SITTIG ³⁵⁶⁾ entnommen werden.

356) HEATH, H. (1952) - "A factor analysis of women's measurements taken for garment and pattern construction", Psychometrika 17, Seite 87-100.
 SITTIG, J. (1953) - "Standardization of garment sizing by statistical methods", Bull. Inst. Int. Stat. 34, 4ème Livr., Seite 307-316.
 - - Ders. (1959) - "Produktiviteitsverhoging op groote schaal: Maatnormalisatie van kleding", Stat. Neerl. 13, Seite 155-169.

VERZEICHNIß DER ABDRUCKUNGEN FÜR ZEITSCHRIFTEN

Allg.Stat.Arch.	Allgemeines Statistisches Archiv
Am.J.Psychol.	American Journal of Psychology
Am.Math.Monthly	American Mathematical Monthly
Am.Soc.Rev.	American Sociological Review
Ann.Eugen.	Annals of Eugenics
Ann.Inst.Stat.Math.	Annals of the Institute of Statistical Mathematics
Ann.Math.Stat.	Annals of Mathematical Statistics
Appl.Stat.	Applied Statistics
Arch.f.d.ges.Psychol.	Archiv für die gesamte Psycholo- gie
Biom.Zeitschr.	Biometrische Zeitschrift
Brit.J.Appl.Phys.	British Journal of Applied Physics
Brit.J.Educ.Psychol.	British Journal of Educational Psychology
Brit.J.Psychol.	British Journal of Psychology (General Section)
Brit.J.Psychol.Stat. Sect.	British Journal of Psychology (Statistical Section) Ab Band 6(1953): Brit.J.Stat. Psychol.
Brit.J.Stat.Psychol.	British Journal of Statistical Psychology Bis Band 5(1952): Brit.J.Psychol. Stat.Sect.
Bull.Inst.Int.Stat.	Bulletin de L'Institut Internatio- nal de Statistique
Bull.Oxf.Univ.Inst. Stat.	Bulletin of the Oxford University Institute of Statistics
Harv.Educ.Rev.	Harvard Educational Review
Ind.Qual.Contr.	Industrial Quality Control
J.Am.Stat.Ass.	Journal of the American Statisti- cal Association
J.Educ.Psychol.	Journal of Educational Psychology
J.Marketing	Journal of Marketing
J.Roy.Stat.Soc.	Journal of the Royal Statistical Society, Series A (General)

J.Roy.Stat.Soc.B.	Journal of the Royal Statistical Society, Series B (Methodological) Bis Band <u>9</u> (1947): J.Roy.Stat.Soc. Suppl.
J.Roy.Stat.Soc.Suppl.	Supplement to the Journal of the Royal Statistical Society Ab Band <u>10</u> (1948): J.Roy.Stat.Soc. B.
Mittbl.Math.Stat.	Mitteilungsblatt für Mathematische Statistik und ihre Anwendungsgebiete
Monthly Bull.Agr.Econ. Stat.	Monthly Bulletin of Agricultural Economics and Statistics
Nachrichtentechn. Fachber.	Nachrichtentechnische Fachberichte
Opn.Res.	Journal of the Operations Research Society of America
Phil.Mag.	The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science
Pop.Stud.	Population Studies
Proc.Amsterdam	Koninklijke Akademie van Wetenschappen, Amsterdam Proceedings of the Section of Science
Proc.Nat.Acad.Sci.USA.	Proceedings of the National Academy of Sciences of the United States of America
Proc.Roy.Soc.Edinb.	Proceedings of the Royal Society of Edinburgh
Proc. 13th Int.Congr. Psychol.Stockholm	Proceedings and Papers, 13th International Congress of Psychology, Stockholm 1951
Psychol.Rev.	Psychological Review
Quart.Appl.Math.	Quarterly of Applied Mathematics
Rev.Econ.Stat.	Review of Economics and Statistics
Roy.Soc.Lond.Phil. Trans.A.	Philosophical Transactions of the Royal Society of London, Series A
Rur.Sociol.	Rural Sociology
Skand.Akt.Tidskr.	Skandinavisk Aktuarietidskrift
Stat.Neerl.	Statistica Neerlandica
Stat.Rundsch.Nordrh.-Westf.	Statistische Rundschau für das Land Nordrhein-Westfalen

Wirtsch.u.Stat.N.F.	Wirtschaft und Statistik, Neue Folge
Zeitschr.d.Bayer.Stat.Landesamts	Zeitschrift des Bayerischen Statistischen Landesamts
Zeitschr.f.Psychol.	Zeitschrift für Psychologie
Zeitschr.f.Schweiz.Stat.u.Volksw.	Zeitschrift für Schweizerische Statistik und Volkswirtschaft

LITERATURVERZEICHNIS

- ADCOCK, C.J. (1954) - Factorial Analysis for Non-Mathematicians. Carlton: Melbourne University Press
- AITKEN, A.C. (1937) - "Studies in practical mathematics. II. The evaluation of the latent roots and latent vectors of a matrix", Proc.Roy.Soc.Edinb. 57, Seite 269-303
- AKAIKE, H. (1958) - "On a computation method for eigenvalue problems and its application to statistical analysis", Ann.Inst.Stat.Math. 10, Seite 1-20
- ALBERT, A.A. (1944a) - "The matrices of factor analysis", Proc.Nat.Acad.Sci.USA. 30, Seite 90-95
- (1944b) - "The minimum rank of a correlation matrix", Proc.Nat.Acad.Sci.USA. 30, Seite 144-146
- AMATO, V. (1958) - Forme Quadratiche in Statistica. Mailand: A.Giuffrè, 151 Seiten
- ANDERSON, O. (1957³) - Probleme der Statistischen Methodenlehre in den Sozialwissenschaften (Einzelschr. d. Deutsch.Stat.Ges.Nr.6). Würzburg: Physica-Verlag, VIII+358 Seiten
- ANDERSON, T.W. (1958) - An Introduction to Multivariate Statistical Analysis. New York: J.Wiley & Sons (London: Chapman & Hall), XII+374 Seiten
- (1959) - Some Scaling Models and Estimation Procedures in the Latent Class Model. In: GRENANDER, U. (Hrsg) - Probability and Statistics. The HARALD CRAMER Volume. Stockholm: Almqvist & Wiksell (New York: J.Wiley & Sons), Seite 9-38
- BABINGTON-SMITH, B. (1950) - "The evaluation of factor analysis from the point of view of a psychologist" (Teil 2 von KENDALL, M.G.; BABINGTON-SMITH, B. - "Factor analysis"), J.Roy.Stat.Soc.B. 12, Seite 73-85

- BALL, R.J.; HAZLEWOOD, A.; KLEIN, L.R. (1959) - "Econometric forecasts for 1959", Bull.Oxf.Univ.Inst.Stat. 21, Seite 3-16
- BANKS, C.; BURT, C. (1947) - "Factor analysis of body measurements for British adult males", Ann.Eugen. 13, Seite 238-256
- BANKS, C. (1954) - "The factorial analysis of crop productivity", J.Roy.Stat.Soc.B. 16, Seite 100-111
- BARAKAT, M.K. (1951) - "A factorial study of mathematical abilities", Brit.J.Psychol.Stat.Sect. 4, Seite 137-156
- BARGER, H. (1942) - Outlay and Income in the United States 1921-1938
- BARGMANN, R. (1956) - "Überführung der Schwerpunkts- in die Hauptachsenlösung und weitere Entwicklung der Näherungsverfahren in der Faktorenanalyse", Mittbl. Math.Stat. 8, Seite 1-14
- BARTELS, H.; FÜRST, G. (1949) - "Preisindices im volkswirtschaftlichen Güterkreislauf", Wirtsch.u.Stat.N.F. 1, Seite 261-267
- BARTLETT, M.S. (1937) - "The statistical conception of mental factors", Brit.J.Psychol. 28, Seite 97-104
- (1947) - "Multivariate analysis", J.Roy.Stat. Soc.Suppl. 2, Seite 176-190
- BASS, B.M. (1957) - "Iterative inverse factor analysis - a rapid method for clustering persons", Psychometrika 22, Seite 105-107
- BERNYER, G. (1957) - "Psychological factors: Their number, nature, and identification", Brit.J.Stat.Psychol. 10, Seite 17-27
- (1958) - "Second order factors and the organization of cognitive functions", Brit.J.Stat.Psychol. 11, Seite 19-29
- BIEBERBACH, L. (1944⁵) - Differentialrechnung (Teubners Mathematische Leitfäden, Bd. 4). Leipzig (Berlin): B.G. Teubner, VI+142 Seiten
- BOGAARD, P.J.M. VAN DEN; KAPTEIN, E. (1958) - Principal Components. Netherlands School of Economics: Econometric Institute, Report 5812, 20 Seiten
- BOGUE, D.J. (1955) - "Nodal versus homogeneous regions and statistical techniques for measuring the influence of each", Bull.Inst.Int.Stat. 35, 4ème Livr., Seite 377-397
- BOUSTEDT, O. (1951) - "Die statistischen Probleme der sozial-ökonomischen Topographie", Allg.Stat.Arch. 35, Seite 291-310
- (1955) - "Statistical methods and problems in regional research", Bull.Inst.Int.Stat. 35, 4ème Livr., Seite 357-375

- BOUSTEDT, O.; RANZ, H. (1957) - Regionale Struktur- und Wirtschaftsforschung (Veröffentl. d. Akad. f. Raumf. u. Landesplg., Abhandl. Bd. 33). Bremen: W. Dorn Verlag, 218 Seiten
- BOWLEY, A. L. (1937⁶) - Elements of Statistics (Studies in Economics and Political Science No. 8). London (New York): Staples Press, XV+503 Seiten
- BRANDNER, H. (1953) - "Der Erzeugerpreisindex industrieller Produkte. Endgültige Berechnung", Wirtsch. u. Stat. N.F. 5, Seite 251-256
- (1958) - "Neuberechnung eines Index der Erzeugerpreise forstwirtschaftlicher Produkte", Wirtsch. u. Stat. N.F. 10, Seite 498-501
- BROWN, W.; THOMSON, G. H. (1925³) - The Essentials of Mental Measurement. Cambridge: At the University Press, X+224 Seiten
- BUCKATZSCH, E. J. (1947) - "The influence of social conditions on mortality rates", Pop. Stud. 1, Seite 229-248
- BURT, C. (1937) - "Correlations between persons", Brit. J. Psychol. 28, Seite 59-96
- (1940) - The Factors of Mind. London: The University of London Press
- (1944) - "The factorial study of physical types", Man 44, Seite 82-86
- (1947) - "Factor analysis and physical types", Psychometrika 12, Seite 171-188
- (1948a) - "A comparison of factor analysis and analysis of variance", Brit. J. Psychol. Stat. Sect. 1, Seite 3-26
- (1948b) - "Factor analysis and canonical correlations", Brit. J. Psychol. Stat. Sect. 1, Seite 95-106
- (1949) - "Alternative methods of factor analysis and their relations to PEARSON'S method of 'principal axes'", Brit. J. Psychol. Stat. Sect. 2, Seite 98-121
- (1951) - "The numerical solution of linear equations", Brit. J. Psychol. Stat. Sect. 4, Seite 31-54
- BURT, C.; WATSON, H. (1951) - "Factor analysis of assessments for a single person", Brit. J. Psychol. Stat. Sect. 4, Seite 179-191
- CARROLL, B. (1953) - "An analytical solution for approximating simple structure in factor analysis", Psychometrika 18, Seite 23-38
- CARTER, L. F.; DUDEK, F. J. (1947) - "The use of psychological techniques in measuring and critically analyzing navigators' flight performance", Psychometrika 12, Seite 31-42

- CATTELL, R.B. (1947) - "Confirmation and clarification of primary personality factors", *Psychometrika* 12, Seite 197-220
- (1948) - "The integration of factor analysis with psychology", *J.Educ.Psychol.* 39, Seite 227-236
- (1949) - "A note on factor invariance and the identification of factors", *Brit.J.Psychol.Stat. Sect.* 2, Seite 134-139
- (1952) - *Factor Analysis: An Introduction and Manual for the Psychologist and Social Scientist*. New York: Harper & Brothers, XIII+462 Seiten
- CHAUNCEY, H. (1939) - "A method for solving simultaneous equations, with particular reference to multiple correlation problems", *Harv.Educ.Rev.* 9, Seite 63-68
- CORSTEN, L.C.A. (1958) - *Vectors, a Tool in Statistical Regression Theory* (Mededelingen van de Landbouwhogeschool te Wageningen, 58/1). Wageningen: Veenman & Zonen, 92 Seiten
- CRAMER, H. (1946) - *Mathematical Methods of Statistics*. Princeton: Princeton University Press, XVI+575 Seiten
- CREAGER, J.A. (1958) - "General resolution of correlation matrices into components and its utilization in multiple and partial regression", *Psychometrika* 23, Seite 1-8
- CREASY, M.A. (1957) - "Analysis of variance as an alternative to factor analysis", *J.Roy.Stat.Soc.B.* 19, Seite 318-325
- CROWTHER, G. (1934) - "The 'Economist' Index of Business Activity", *J.Roy.Stat.Soc.* 97, Seite 241-261
- CURETON, E.E. (1939) - "The principal compulsions of factor-analysts", *Harv.Educ.Rev.* 9, Seite 287-295
- DAMM, H. (1914) - *Korrelative Beziehungen zwischen elementaren Vergleichsleistungen* (Zeitschr.f. angew. Psychol. u. psychol. Sammlf., Beih. 9). Leipzig: J.A. Barth, IV+87 Seiten
- DAVIS, H.T. (1941) - *The Analysis of Economic Time Series* (COWLES Commission for Research in Economics Monogr. No. 6). Bloomington: The Principia Press, XIV+620 Seiten
- DENEFFE, P. (1949) - "Soziale Topographie der Großstadt", *Allg.Stat.Arch.* 32, Seite 444-471
- DENEFFE, P.; BRANDNER, H. (1952) - "Erzeugerpreise landwirtschaftlicher Produkte", *Wirtsch.u.Stat.N.F.* 4, Seite 61-65

- DENEFFE, P.; KELLER (1952) - "Zur Berechnung eines Index der Einzelhandelspreise", Wirtsch.u.Stat.N.F. 4 Seite 363-369
- DODD, S.C. (1928a) - "The theory of factors, I", Psychol.Rev. 35, Seite 211-234
- (1928b) - "The theory of factors, II", Psychol.Rev. 35, Seite 261-279.
- DUBOIS, P.H. (1957) - Multivariate Correlational Analysis. New York: Harper & Brothers, XV+202 Seiten
- DUSCHEK, A. (1949-50) - Vorlesungen über höhere Mathematik. Zwei Bände. Wien: Springer-Verlag, X+395 und VI+386 Seiten
- DWYER, P.S. (1940) - "The evaluation of multiple and partial correlation coefficients from the factorial matrix", Psychometrika 5, Seite 211-232
- (1941a) - "The evaluation of determinants", Psychometrika 6, Seite 191-204
- (1941b) - "The solution of simultaneous equations", Psychometrika 6, Seite 101-129
- (1951) - Linear Computations. New York: J.Wiley & Sons (London: Chapman & Hall), XI+344 Seiten
- ECKART, C.; YOUNG, G. (1936) - "The approximation of one matrix by another of lower rank", Psychometrika 1, Seite 211-218
- EDGEWORTH, F.Y. (1896) - "Supplementary notes on statistics", J.Roy.Stat.Soc. 59, Seite 529-539
- EGERT, P. (1958) - Kraftfahrzeugbestand und Verkehrsunfälle (Verkehrswiss.Veröffentl.d.Min.f.Wirtsch.u.Verkehr Nordrh.-Westf., Heft 40). Düsseldorf: Droste-Verlag, 76 Seiten
- EIJK, C.J. VAN (1957) - "Economische toepassingen van factoranalyse", Stat.Neerl. 11, Seite 211-213
- ENGELMANN, M. (1960) - "Der Verbrauch in Arbeitnehmerhaushalten im Jahr 1959", Wirtsch.u.Stat.N.F. 12, Seite 207-211
- ESENWEIN-ROTHE, I. (1955) - "Statistik als Methode der Raumforschung", Allg.Stat.Arch. 39, Seite 136-143
- EYSENCK, H.J. (1952) - "Uses and abuses of factor analysis", Appl.Stat. 1, Seite 45-49
- EZEKIEL, M.; FOX, K.A. (1959³) - Methods of Correlation and Regression Analysis. New York: J.Wiley & Sons (London: Chapman & Hall), XV+548 Seiten
- FEIGL, G. (Hrsg. ROHRBACH, H.) (1953) - Einführung in die höhere Mathematik. Berlin (Göttingen, Heidelberg): Springer-Verlag, VIII+375 Seiten

- FERGUSON, G.A. (1954) - "The concept of parsimony in factor analysis", Psychometrika 19, Seite 281-290
- FLOOD, M.M. (1940) - "A computational procedure for the method of principal components", Psychometrika 5, Seite 169-172
- FOX, L. (1950) - "Practical methods for the solution of linear equations and the inversion of matrices", J.Roy.Stat.Soc.B. 12, Seite 120-136
- FRISCH, R. (1934) - Statistical Confluence Analysis by Means of Complete Regression Systems. Oslo: Universitets Økonomiske Institutt, Publikasjon nr. 5, 192 Seiten
- FRUCHTER, B. (1954) - Introduction to Factor Analysis. New York: D. Van Nostrand Co., XII+280 Seiten
- FÜRST, G.; DENEFFE, P. (1952) - "Der neue Preisindex für die Lebenshaltung", Wirtsch.u.Stat.N.F. 4, Seite 439-455
- FÜRST, G. (1955) - "Gedanken zur regionalen Gliederung in der Wirtschaftsstatistik", Wirtsch.u.Stat.N.F. 7, Seite 181-187
- (1960) - "Zur Aussagekraft von Preisindexziffern für die Lebenshaltung", Wirtsch.u.Stat.N.F. 12, Seite 5-8
- GANTMACHER, F.R. (1958) - Matrizenrechnung. Teil I: Allgemeine Theorie (Hochschulbücher f. Mathematik, Bd. 36). Berlin: VEB Deutscher Verlag der Wissenschaften, XI+324 Seiten
- GEARY, B. (1948) - "Studies in relations to economic time series", J.Roy.Stat.Soc.B. 10, Seite 140-158
- GIBSON, W.A. (1952) - "Orthogonal and oblique simple structures", Psychometrika 17, Seite 317-323
- GIRSHICK, M.A. (1936) - "Principal components", J.Am.Stat.Ass. 31, Seite 519-528
- (1939) - "On the sampling theory of the roots of determinantal equations", Ann.Math.Stat. 10, Seite 203-224
- GLICKMAN, A.S. (1958) - "Factor analysis of personnel components of ship performance", Opn.Res. 6, Seite 106-115
- GOSNELL, H.F.; SCHMIDT, M.J. (1936) - "Factorial and correlational analysis of the 1934 vote in Chicago", J.Am.Stat.Ass. 31, Seite 507-518
- GOTTSCHALDT, K. (1940) - Die Statistik in der Psychologie. In: Die Statistik in Deutschland nach ihrem heutigen Stand. Ehrengabe für FRIEDRICH ZAHN. Hrsg. von F. BURGDÖRFER. Berlin: Verlag f. Sozialpol., Wirtsch.u.Stat. P. Schmidt, Seite 386-396

- GREEN, B.F. (1952) - "Latent structure analysis and its relation to factor analysis", J.Am.Stat.Ass. 47, Seite 71-76
- GRIESMEIER, J. (1955) - "Berechnung wirtschaftlicher Leistungswerte für kleinere Gebietseinheiten", Allg. Stat.Arch. 39, Seite 17-24
- GROTEMEYER, K.P. (1958) - Analytische Geometrie (Sammlung Götschen, Bd. 65/65a). Berlin: Walter de Gruyter & Co, 202 Seiten
- GRÜNEBERG, H.J. (1952) - "Die multiple Faktoranalyse", Mittbl. Math.Stat. 4, Seite 9-30
- GUILFORD, J.P. (1952) - "Factor analysis of the Army Air Forces SHEPPARD Field Battery of experimental aptitude tests", Psychometrika 17, Seite 45-63
- GUTTMAN, L. (1954a) - "Some necessary conditions for common-factor analysis", Psychometrika 19, Seite 149-161
- (1954b) - A New Approach to Factor Analysis: The Radex. In: LAZARSFELD, P.F. (Hrsg.) - Mathematical Thinking in the Social Sciences. Glencoe: The Free Press, Seite 349-387
- (1955) - "A generalized simplex for factor analysis", Psychometrika 20, Seite 173-192
- (1957) - "Simple proofs of relations between the communality problem and multiple correlation", Psychometrika 22, Seite 147-157
- (1958) - "To what extent can communalities reduce rank ?", Psychometrika 23, Seite 297-308
- HAGOOD, M.J.; DANILEVSKY, N.; BEUM, C.O. (1941) - "An examination of the use of factor analysis in the problem of subregional delineation", Rur.Sociol. 6, Seite 216-233
- HAGOOD, M.J. (1943) - "Statistical methods for delineation of regions applied to data on agriculture and population", Social Forces 21, Seite 287-297
- HAGOOD, M.J.; BERNERT, E.H. (1945) - "Component indexes as a basis for stratification in sampling", J.Am.Stat.Ass. 40, Seite 330-341
- HAGOOD, M.J.; PRICE, D.O. (1952²) - Statistics for Sociologists. New York: H.Holt & Co, 575 Seiten
- HALL, D.M.; WELKER, E.L.; CRAWFORD, I. (1945) - "Factor analysis calculations by tabulating machines" Psychometrika 10, Seite 93-125
- HAMILTON, M. (1958) - "An experimental approach to the identification of factors", Brit.J.Stat.Psychol. 11, Seite 159-169
- HAMMOND, W.H. (1942) - "An application of BURT'S multiple general factor analysis to the delineation of physical types", Man 42, Seite 4-11

- HAMMOND, W.H. (1946) - "Factor analysis as applied to social and economic data", Brit.J.Educ.Psychol. 16, Seite 178
- HARPER, R.; BARON, M. (1948) - "Factorial analysis of the rheological properties of cheese", Nature 162, Seite 821
- HARPER, R.; KENT, A.J.; BLAIR, G.W.S. (1950) - "The application of multiple factor analysis to industrial test data", Brit.J.Appl.Phys. 1, Seite 23-28
- HARPER, R. (1956) - "Factor analysis as a technique for examining complex data on foodstuffs", Appl.Stat. 2, Seite 32-48
- HARRIS, C.W.; KNOELL, D.M. (1948) - "The oblique solution in factor analysis", J.Educ.Psychol. 39, Seite 385-403
- HEATH, H. (1952) - "A factor analysis of women's measurements taken for garment and pattern construction", Psychometrika 17, Seite 87-100
- HENRYSSON, S. (1957) - Applicability of Factor Analysis in the Behavioral Sciences. A Methodological Study (Stockholm Studies in Educational Psychology I). Stockholm: Almqvist & Wiksell, 156 Seiten
- HOEL, P.G. (1957²) - Introduction to Mathematical Statistics. New York: J.Wiley & Sons (London: Chapman & Hall), XI+331 Seiten
- HOFSTÄTTER, P.R. (1938) - "Über Faktoren-Analyse", Arch.f.d. ges.Psychol. 100, Seite 223-279
- (1953) - Einführung in die quantitativen Methoden der Psychologie. München: J.A.Barth, IV+175 Seiten
- HOLZINGER, K.J.; SWINEFORD, F. (1937) - "The bi-factor method", Psychometrika 2, Seite 41-54
- HOLZINGER, K.J.; HARMAN, H.H. (1941) - Factor Analysis. A Synthesis of Factorial Methods. Chicago: The University of Chicago Press, XII+417 Seiten
- HOLZINGER, K.J. (1946) - "A comparison of the principal-axes and centroid factors", J.Educ.Psychol. 37, Seite 449-472
- HOOPER, J.W. (1959) - "Simultaneous equations and canonical correlation theory", Econometrica 27, Seite 245-256
- HORST, P. (1937) - "A method of factor analysis by means of which all coordinates of the factor matrix are given simultaneously", Psychometrika 2, Seite 225-236
- (1941) - "A non-graphical method for transforming an arbitrary factor matrix into a simple structure factor matrix", Psychometrika 6, Seite 79-99

- HOSSE, H. (1957) - "Zur Problematik der deutschen Straßenverkehrsunfallstatistik", Allg. Stat. Arch. 41, Seite 18-32
- HOTELLING, H. (1931) - "The generalization of STUDENT'S ratio", Ann. Math. Stat. 2, Seite 360-378
- (1933) - "Analysis of a complex of statistical variables into principal components", J. Educ. Psychol. 24, Seiten 417-441 und 498-520
- (1936a) - "Simplified calculation of principal components", Psychometrika 1, Seite 27-35
- (1936b) - "Relations between two sets of variates", Biometrika 28, Seite 321-377
- (1943) - "Some new methods in matrix calculation", Ann. Math. Stat. 14, Seite 1-34
- (1947) - Multivariate Quality Control, Illustrated by the Air Testing of Sample Bombsights. In: EISENHART, C. (Hrsg.) - Selected Techniques of Statistical Analysis for Scientific and Industrial Research and Production and Management Engineering. New York (London): McGraw Hill Book Co, Seite 113-184
- (1957) - "The relations of newer multivariate statistical methods to factor analysis", Brit. J. Stat. Psychol. 10, Seite 67-79
- HOUSEHOLDER, A.S.; YOUNG, G. (1938) - "Matrix approximation and latent roots", Am. Math. Monthly 45, Seite 165-171
- HOUSEHOLDER, A.S. (1956) - "Numerical mathematics from the viewpoint of electronic digital computers", Nachrichtentechn. Fachber. 4, Seite 21-25
- HSU, P.L. (1939) - "On the distribution of roots of certain determinantal equations", Ann. Eugen. 9, Seite 250-258
- JACKSON, J.E. (1956) - "Quality control methods for two related variables", Ind. Qual. Contr. 12, No. 7, Seite 4-8
- JACKSON, J.E.; MORRIS, R.H. (1957) - "An application of multivariate quality control to photographic processing", J. Am. Stat. Ass. 52, Seite 186-199
- JACKSON, J.E. (1959) - "Quality control methods for several related variables", Technometrics 1, Seite 369-377
- JACOBS, A. (1957) - Der internationale Vergleich der Lebenshaltungskosten. Eine methodologische Untersuchung dargestellt an Deutschland-Frankreich. Berlin: Duncker & Humblot, 117 Seiten
- JANOSI, P.E. DE (1959) - "Factors influencing the demand for new automobiles", J. Marketing 23, Seite 412-418

- JOWETT, G.H. (1958) - "Factor analysis", Appl. Stat. 7,
Seite 114-125
- KELLERER, H. (1960) - Statistik im modernen Wirtschafts- und
Sozialleben (rowohlts deutsche enzyklopädie, Bd.
103/104). Hamburg: Rowohlt, 289 Seiten
- KELLEY, T.L. (1935) - Essential Traits of Mental Life (Har-
vard Studies in Education, Vol. 26). Cambridge:
Harvard University Press
- (1940) - "Comment on WILSON and WORCESTER'S
'Note on factor analysis'", Psychometrika 5, Seite
117-120
- KELLOG, C.E. (1936a) - "The problem of principal components:
Derivation of HOTELLING'S method from THURSTONE'S",
J. Educ. Psychol. 27, Seite 512-520
- (1936b) - "The problem of principal components:
II - The argument for communalities", J. Educ. Psychol. 27,
Seite 581-590
- KENDALL, M.G. (1939) - "The geographical distribution of crop
productivity in England", J. Roy. Stat. Soc. 102,
Seite 21-48
- (1948) - Rank Correlation Methods. London:
Charles Griffin & Co, VII+160 Seiten
- (1950) - "Factor analysis as a statistical tech-
nique" (Teil 1 von KENDALL, M.G.; BABINGTON-SMITH, B.
- "Factor analysis"), J. Roy. Stat. Soc. B. 12, Seite
60-73
- (1957) - A Course in Multivariate Analysis
(Griffin's Statistical Monographs and Courses, No. 2).
London: Charles Griffin & Co, 185 Seiten
- KENDALL, M.G.; BUCKLAND, W.R. (1957) - A Dictionary of Sta-
tistical Terms. London (Edinburgh): Oliver & Boyd,
IX+493 Seiten
- KINCAID, W.M. (1947) - "numerical methods for finding cha-
racteristic roots and vectors of matrices", Quart.
Appl. Math. 5, Seite 320-345
- KLAYMAN, M.J. (1960) - "International index numbers of food
and agricultural production", Monthly Bull. Agr.
Econ. Stat. 9, Seite 12-14
- KLEIN, L.R.; GOLDBERGER, A.S. (1955) - An Econometric Model
of the United States 1929-1952 (Contributions to
Economic Analysis IX). Amsterdam: North-Holland
Publishing Co, XV+165 Seiten
- KLEINE, W. (1959) - "Wetter und Straßenverkehrsunfälle in
Nordrhein-Westfalen", Stat. Rundsch. Nordrh.-Westf.
11, Seite 217-223
- (1960) - "Straßen-, Kraftfahrzeug- und Unfall-
dichte in Nordrhein-Westfalen im Jahre 1958", Stat.
Rundsch. Nordrh.-Westf. 12, Seite 32-36

- KLINGBERG, F.L. (1941) - "Studies in measurement of the relations among sovereign states", Psychometrika 6, Seite 335-352
- KLOEK, T.; MENNES, L.B.M. (1960) - "Simultaneous equations estimation based on principal components of pre-determined variables", Econometrica 28, Seite 45-61
- KRUEGER, F.; SPEARMAN, C. (1906) - "Die Korrelation zwischen verschiedenen geistigen Leistungsfähigkeiten", Zeitschr.f.Psychol. 44, Seite 50-114
- KUZNETS, S. (1941) - National Income and its Composition, 1919-1938, Volume I. New York: National Bureau of Economic Research, XXX+929 Seiten
- LANSING, J.B.; BLOOD, D.M. (1958) - "A cross-section analysis of non-business air travel", J.Am.Stat.Ass. 53, Seite 928-947
- LAWLEY, D.N. (1956) - "Tests of significance for the latent roots of covariance and correlation matrices", Biometrika 43, Seite 128-136
- LAZARSFELD, P.F. (1929) - Statistisches Praktikum für Psychologen und Mediziner. Jena: G.Fischer, VIII+180 Seiten
- LAZARSFELD, P.F. (1954) - A Conceptual Introduction to Latent Structure Analysis. In: LAZARSFELD, P.F. (Hrsg.) - Mathematical Thinking in the Social Sciences. Glencoe: The Free Press, Seite 349-387
- LEDERMANN, W. (1937) - "On the rank of the reduced correlational matrix in multiple factor analysis", Psychometrika 2, Seite 85-93
- (1940) - "A problem concerning matrices with variable diagonal elements", Proc.Roy.Soc.Edinb. 60, Seite 1-17
- LEVE, G. DE (1957) - "Enige statistische aspecten van de factoranalyse", Stat.Neerl. 11, Seite 201-209
- LIENERT, G.A. (1959) - "Prinzip und Methode der multiplen Faktorenanalyse, demonstriert an einem Beispiel", Biom.Zeitschr. 1, Seite 88-141
- LINDE, H. (1952) - "Grundfragen der Gemeindetypisierung", Forschungs-u.Sitzungsber.d.Akad.f.Raumf.u.Landesplg. 2, Seite 58-121
- LIVELY, C.E.; ALMACK, R.B. (1938) - A Method of Determining Rural Social Sub-Areas with Application to Ohio (Ohio State University and Ohio Agricultural Experiment Station, Department of Rural Economics Mimeograph Bulletin No. 106).
- LOCKE, H.J.; WILLIAMSON, R.C. (1958) - "Marital adjustment: A factor analysis study", Am.Soc.Rev. 23, Seite 562-569

- MANGOLDT, H.v. (Bearb. KNOPP, K.) (1948⁹) - Einführung in die höhere Mathematik. Drei Bände. Stuttgart: S. Hirzel Verlag, XV+585, XV+634 und XVI+618 Seiten
- MCCLOY, C.H.; METHENY, E.; KNOTT, V. (1938) - "A comparison of the THURSTONE method of multiple factors with the HOTELLING method of principal components", Psychometrika 3, Seite 61-67
- MCNEMAR, Q. (1942) - "On the number of factors", Psychometrika 7, Seite 9-18
- MEERWARTH, R. (1920) - Einleitung in die Wirtschaftsstatistik. Jena: G. Fischer, VI+329 Seiten
- (1925) - Nationalökonomie und Statistik (Handb. d. Wirtsch.-u. Sozialwiss. Bd. 7). Berlin (Leipzig): Walter de Gruyter & Co, XII+506 Seiten
- MILLICAN, R.D. (1959) - "Application of factor analysis to consumption expenditures", Rev. Econ. Stat. 41, Seite 74-76
- MINISTERIUM FÜR WIRTSCHAFT UND VERKEHR, NORDRHEIN-WESTFALEN (Hrsg.) (1956) - Die Verkehrsentwicklung des Landes Nordrhein-Westfalen, Nr. 27: Jahresbericht 1955, 193 Seiten
- NEISS, F. (1955⁴) - Determinanten und Matrizen. Berlin (Göttingen, Heidelberg): Springer-Verlag, 111 Seiten
- NEUHAUS, J.O.; WRIGLEY, C. (1954) - "The quartimax method. An analytic approach to orthogonal simple structure", Brit. J. Stat. Psychol. 7, Seite 81-91
- ORGANISATION FOR EUROPEAN ECONOMIC CO-OPERATION (Hrsg.) (1955) - The Measurement of Agricultural Production and Food Consumption. 41 Seiten
- PEARSON, K. (1896) - "Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia", Roy. Soc. Lond. Trans. A. 187, Seite 253-318
- (1901) - "On lines and planes of closest fit to systems of points in space", Phil. Mag. 6th Ser. 2, Seite 579-572
- PEEL, E.A. (1953) - Factorial Analysis as a Psychological Technique. In: Uppsala Symposium on Psychological Factor Analysis (Nordisk Psykologi's Monograph Series No. 3). Stockholm: Almqvist & Wiksell, Seite 7-22
- PETERS, C.C.; VAN VOORHIS, W.R. (1940) - Statistical Procedures and their Mathematical Bases. New York (London): McGraw Hill Book Co, XI+516 Seiten

- FFANZAGL, J. (1960) - Allgemeine Methodenlehre der Statistik I. Elementare Methoden unter besonderer Berücksichtigung der Anwendungen in den Wirtschafts- und Sozialwissenschaften (Sammlung Göchen, Bd. 746/746a). Berlin: Walter de Gruyter & Co, 205 Seiten
- PILLAI, K.C.S. (1956) - "On the distribution of the largest or smallest root of a matrix in multivariate analysis", *Biometrika* 43, Seite 122-127
- POTI, S.J. (1946) - "A study of expenditure patterns of Calcutta middle class family budgets", *Sankhyā* 7, Seite 425-428
- PRICE, D.O. (1942) - "Factor analysis in the study of metropolitan centers", *Social Forces* 20, Seite 449-455
- QUENOUILLE, M.H. (1952) - Associated Measurements. London: Butterworths Scientific Publications, X+242 Seiten
- RAO, C.R. (1949) - "Representation of 'p' dimensional data in lower dimensions" ("Anthropometric Survey of the United Provinces, 1941: A Statistical Study" Appendix 4), *Sankhyā* 9, Seite 248-251
- (1952) - Advanced Statistical Methods in Biometric Research. New York: J. Wiley & Sons (London: Chapman & Hall), 390 Seiten
- (1955) - "Estimation and tests of significance in factor analysis", *Psychometrika* 20, Seite 93-111
- REIERSØL, O. (1940) - "A method for recurrent computation of all the principal minors of a determinant and its application in confluence analysis", *Ann. Math. Stat.* 11, Seite 193-198
- REYBURN, H.A.; TAYLOR, J.G. (1943) - "On the interpretation of common factors: A criticism and a statement", *Psychometrika* 8, Seite 53-64
- REYBURN, H.A.; RAATH, M.J. (1949) - "Simple structure: A critical examination", *Brit. J. Psychol. Stat. Sect.* 2, Seite 125-133
- RHODES, E.C. (1927) - "On lines and planes of closest fit", *Phil. Mag.* 7th Ser. 3, Seite 357-364
- (1937) - "The construction of an Index of Business Activity", *J. Roy. Stat. Soc.* 100, Seite 18-39
- RICHARDSON, L.F. (1950a) - "A purification method for computing the latent roots of numerical matrices and some integrals of differential equations", *Roy. Soc. Lond. Phil. Trans. A.* 242, Seite 439-491
- (1950b) - "A method for computing principal axes", *Brit. J. Psychol. Stat. Sect.* 3, Seite 16-20

- ROPPERT, J. (1959) - Einige Anwendungen des Matrizenkalküls bei der Untersuchung von statistischen Verteilungen. In: Anwendungen der Matrizenrechnung auf wirtschaftliche und statistische Probleme (Einzelschr. d. Deutsch. Stat. Ges. Nr. 9). Würzburg: Physica-Verlag, Seite 216-234
- ROSNER, B. (1948) - "An algebraic solution for the communalities", Psychometrika 13, Seite 181-184
- RÜDIGER, W. (1956) - "Neuberechnung des Index der industriellen Nettoproduktion", Wirtsch. u. Stat. N. F. 8, Seite 129-146
- SANAI, M. (1952) - "An empirical study of political, religious, and social attitudes", Brit. J. Psychol. Stat. Sect. 5, Seite 81-91
- SCHAAFSMA, A. H.; WILLEMZE, F. G. (1959²) - Moderne Qualitätskontrolle. Eindhoven: Philips Technische Bibliothek, XII+493 Seiten
- SCHAUB, G. (1958) - "Zur Problematik der deutschen Straßenverkehrsunfallstatistik. Erwiderung und Ergänzung zum Aufsatz von HOSSE", Allg. Stat. Arch. 42, Seite 29-38
- SCHOLZ, H. (1959) - Praktische Lösung von Aufgaben der Matrizenrechnung. In: Anwendungen der Matrizenrechnung auf wirtschaftliche und statistische Probleme (Einzelschr. d. Deutsch. Stat. Ges. Nr. 9). Würzburg: Physica-Verlag, Seite 235-262
- SCHUESSLER, K. F.; DRIVER, H. (1956) - "A factor analysis of sixteen primitive societies", Am. Soc. Rev. 21, Seite 492-499
- SHARMA, O. C. (1954) - "Factor analysis of technical trades and educational examination marks of the aircraftsmen of the Indian Air Force", Sankhyā 13, Seite 27-34
- SITTIG, J. (1953) - "Standardization of garment sizing by statistical methods", Bull. Inst. Int. Stat. 34, 4ème Livr., Seite 307-316
- (1959) - "Produktiviteitsverhoging op grote schaal: Maatnormalisatie van kleding", Stat. Neerl. 13, Seite 155-169
- SPEARMAN, C. (1904a) - "The proof and measurement of association between two things", Am. J. Psychol. 15, Seite 72-101
- (1904b) - "'General Intelligence', objectively determined and measured", Am. J. Psychol. 15, Seite 201-293

- SPERNER, E. (1957^{3,2}) - Einführung in die Analytische Geometrie und Algebra. Zwei Teile (Studia Mathematica, Bde. I u. VI). Göttingen: Vandenhoeck & Ruprecht, VIII+339 und 388 Seiten
- STAEBLE, H. (1934) - International Comparison of Food Costs. In: INTERNATIONAL LABOUR OFFICE (Hrsg.) - International Comparisons of Cost of Living (Studies and Reports, Ser. N, No. 20), Seite 1-105
- STATISTISCHES BUNDESAMT (Hrsg.) - Die Industrie der Bundesrepublik Deutschland. Reihe 2: Produktion ausgewählter Erzeugnisse, Produktionsindex, Februar 1960. Wiesbaden: W. Kohlhammer, 28 Seiten
- - Statistischer Bericht, Arb.-Nr. VI/3/123 vom 29. März 1960, 12 Seiten
- STATISTISCHES LANDESAMT NORDRHEIN-WESTFALEN (Hrsg.) (1959) - Die Gemeindestraßen in Nordrhein-Westfalen am 31. März 1956 (Beitr. z. Stat. Nordrh.-Westf., Heft 108), 115 Seiten
- (1960) - Straßenverkehrsunfälle in Nordrhein-Westfalen 1956-1958 (Beitr. z. Stat. Nordrh.-Westf., Heft 119), 76 Seiten
- STEPHENSON, W. (1935) - "Technique of factor analysis", Nature 136, Seite 297
- (1936) - "Introduction to inverted factor analysis, with some applications to studies in orexis", J. Educ. Psychol. 27, Seite 353-367
- STERN, W. (1911) - Die differentielle Psychologie in ihren methodischen Grundlagen. Leipzig: J. A. Barth, IX+503 Seiten
- STONE, R. (1945) - "The analysis of market demand", J. Roy. Stat. Soc. 108, Seite 286-382
- (1947) - "On the interdependence of blocks of transactions", J. Roy. Stat. Soc. Suppl. 9, Seite 1-32
- (1956) - Quantity and Price Indexes in National Accounts. Paris: Organisation for European Economic Co-Operation, 120 Seiten
- SUTHERLAND, J. (1951) - "Factor rotation by the method of extended vectors", Brit. J. Psychol. Stat. Sect. 4, Seite 21-30
- TAYLOR, C. W. (1947) - "A factorial study of fluency in writing", Psychometrika 12, Seite 239-262
- TECOZ, H. F. (1944) - "Utilité des méthodes statistiques dans quelques applications de la psychologie (Données nouvelles)", Zeitschr. f. Schweiz. Stat. u. Volksw. 80, Seite 48-77

- THEIL, H.; KLOEK, T. (1959) - "The statistics of systems of simultaneous economic relationships", Stat. Neerl. 13, Seite 65-89
- THOMSON, G.H. (1947) - "The maximum correlation of two weighted batteries (HOTELLING'S 'Most predictable criterion')", Brit. J. Psychol. Stat. Sect. 1, Seite 27-34
- (1951⁵a) - The Factorial Analysis of Human Ability. London: The University of London Press
- (1951b) - "Factor analysis, its hopes and dangers", Proc. 13th Int. Congr. Psychol. Stockholm, Seite 50-61
- THRALL, R.M.; TORNHEIM, L. (1957) - Vector Spaces and Matrices. New York: J. Wiley & Sons, XII+318 Seiten
- THURSTONE, L.L. (1940) - "Experimental study of simple structure", Psychometrika 5, Seite 153-168
- (1944) - "Graphical method of factoring the correlation matrix", Proc. Nat. Acad. Sci. USA. 30, Seite 129-134
- (1947) - Multiple-Factor Analysis. Chicago: The University of Chicago Press, XIX+535 Seiten
- (1954) - "An analytical method for simple structure", Psychometrika 19, Seite 173-182
- TINTNER, G. (1940) - "The analysis of economic time series", J. Am. Stat. Ass. 35, Seite 93-100
- (1946) - "Some applications of multivariate analysis to economic data", J. Am. Stat. Ass. 41, Seite 472-500
- (1950) - "Some formal relations in multivariate analysis", J. Roy. Stat. Soc. B. 12, Seite 95-101
- (1952) - Econometrics. New York: J. Wiley & Sons, (London: Chapman & Hall), XII+370 Seiten
- (1956) - "Statistik und Ökonometrie", Stat. Vierteljahresschr. 2, Seite 92-100
- TOMAN, S. (1949) - "Die Faktorenanalyse in der Psychologie", Stat. Vierteljahresschr. 2, Seite 57-68
- TUCKER, L.R. (1944a) - "The determination of successive principal components without computation of tables of residual correlation coefficients", Psychometrika 9, Seite 149-153
- (1944b) - "A semi-analytical method of factorial rotation to simple structure", Psychometrika 9, Seite 43-68
- (1955) - "The objective definition of simple structure in linear factor analysis", Psychometrika 20, Seite 209-225

- UNITED NATIONS(Hrsg.)(1954) - Report on International Definition and Measurement of Standards and Levels of Living (Doc. E/CN.3/179;E/CN.5/299), XII+95 Seiten
- UNITED NATIONS, ECONOMIC AND SOCIAL COUNCIL, STATISTICAL COMMISSION(Hrsg.)(1957) - A System of Price and Quantity Indexes for National Accounts (Doc. E/CN.3/L.46), 117 Seiten
- (1958) - General-Purpose Index
Numbers of Retail Prices (Doc. E/CN.3/247), 13 Seiten
- (1960) - Problems and Methods in the Gathering of Representative and Comparable Wholesale Price Series (Doc. E/CN.3/264), 51 Seiten
- UNITED NATIONS, STATISTICAL OFFICE(Hrsg.)(1958) - Methods Used in Compiling the United Nations Price Indexes for Basic Commodities in International Trade (Statistical Papers, Ser.M,No.29), III+66 Seiten
- UVEN,M.J. VAN(1930) - "Adjustment of N points of n -dimensional space to the best linear $(n-1)$ -dimensional space", Proc.Amsterdam 33, Seiten 143-157 und 307-326
- VERNON,P.E.(1946) - "Statistical methods in the selection of Navy and Army Personnel", J.Roy.Stat.Soc.Suppl. 8, Seite 139-153
- VINCENT,D.F.(1953) - "The origin and development of factor analysis", Appl.Stat. 2, Seite 107-117
- VINING,R.(1953) - "Delimitation of economic areas: Statistical conceptions of the spatial structure of an economic system", J.Am.Stat.Ass. 48, Seite 44-64
- WAGENFÜHR,R.(1958) - "Probleme des internationalen Vergleichs des Lebensniveaus", Allg.Stat.Arch. 42, Seite 305-309
- WERNER,K.(1958) - Die Industriestatistik der Bundesrepublik Deutschland. Berlin: Duncker & Humblot, 171 Seiten
- WHIPPLE,H.(1926) - "On the best linear relation connecting three variables", Phil.Mag. 7th Ser. 1, Seite 378-384
- WHITTLE,P.(1953) - "On principal components and least squares methods of factor analysis", Skand.Akt. Tidskr. 35, Seite 223-239
- WILKS,S.S.(1938) - "Weighting systems for linear functions of correlated variables when there is no independent variable", Psychometrika 3, Seite 23-43
- (1950) - Mathematical Statistics. Princeton: Princeton University Press, XI+284 Seiten

- WIRNSHOFER, J. (1958) - "Berechnung wirtschaftlicher Leistungswerte für kreisfreie Städte und Landkreise. Methodische Untersuchung auf Grund von Probeerhebungen für ausgewählte Kreise", Zeitschr.d.Bayer. Stat.Landesamts 90, Seite 1-13
- WOLD, H.; JURÉEN, L. (1953) - Demand Analysis. A Study in Econometrics. New York: J.Wiley & Sons (London: Chapman & Hall), XVI+353 Seiten
- WOLD, H. (1957) - Rezension von KENDALL-BUCKLAND (1957) - A Dictionary of Statistical Terms, Kyklos 10, Seite 345-347
- WOOLF, B. (1951) - "Computation and interpretation of multiple regressions", J.Roy.Stat.Soc.B. 13, Seite 100-119
- WRIGLEY, C.R. (1959) - "The effect upon the communalities of changing the estimate of the number of factors", Brit.J.Stat.Psychol. 12, Seite 35-51
- YULE, G.U.; KENDALL, M.G. (1950¹⁴) - An Introduction to the Theory of Statistics. London: Charles Griffin & Co, XXIV+701 Seiten
- ZIMMERMAN, W.S. (1946) - "A simple graphical method for orthogonal rotation of axes", Psychometrika 11, Seite 51-55
- ZURMÜHL, R. (1957²) - Praktische Mathematik für Ingenieure und Physiker. Berlin (Göttingen, Heidelberg): Springer-Verlag, XI+524 Seiten
- (1958²) - Matrizen. Eine Darstellung für Ingenieure. Berlin (Göttingen, Heidelberg): Springer-Verlag, XV+467 Seiten

LEBENS LAUF

Am 12. November 1929 wurde ich als Sohn des Dr.med. Joachim Wulsten und seiner Ehefrau Charlotte, geb. Redes, in Karlsruhe (Baden) geboren. Ich besuchte die Volksschule in Reichenau (Kr. Zittau) und in Berlin-Zehlendorf, seit Ostern 1939 die Oberschule in Kleinmachnow (Kr. Teltow). Nach einigen kriegsbedingten Unterbrechungen des Schulbesuchs legte ich am 10. Juni 1948 an der Schadow-Schule in Berlin-Zehlendorf die Reifeprüfung ab.

Von 1948 bis 1951 war ich als Schulamtsbewerber (Schulhelfer) an mehreren Grundschulen in Berlin-Schöneberg tätig, daran anschließend bis 1956 als Dolmetscher und Personalsachbearbeiter bei einer amerikanischen Dienststelle in Berlin-Lichterfelde. 1952 begann ich mit dem Studium der Volkswirtschaft an der Freien Universität Berlin, das ich am 15. Februar 1957 mit der Diplomprüfung für Volkswirte abschloß. Seit dem 1. März 1957 bin ich Verwalter einer wissenschaftlichen Assistentenstelle am Institut für Statistik und ihre Anwendungen in den Wirtschafts- und Sozialwissenschaften der Universität München.

